. . . WRITE THE SCRIPTS NEEDED IN TIMES OF WAR - DURING TIMES OF PEACE!

. . .

BENJAMIN K. HENCKEL

NIELS BOHR INSTITUTE

CLEANING AND TRAINING

DATA-DRIVEN PHOTON PARTICLE IDENTIFICATION IN THE ATLAS EXPERIMENT

BENJAMIN K. HENCKEL

PUBLISHED BY BENJAMIN K. HENCKEL

TUFTE-LATEX.GITHUB.IO/TUFTE-LATEX/

First printing, December 2019

Contents

List of Figures

List of Tables

Abstract

The current photon identification deployed in the ATLAS detector use rectangular cuts, while this method is performing acceptably at the current environment of the Large Hadron Collider, imminent changes will create an environment, where the current method will be put under severe pressure. It is therefore vital that significant improvements are made. The obvious place to look for improvements are in the use machine learning methods and that is exactly what this work has done. It has trained boosted decision trees and to an extent neural networks, on both Monte Carlo and data datasets. The Monte Carlo trained models have, both in Monte Carlo and data, shown consistent and significant improvements when compared to the current method. The data-driven models show promising preliminary results when evaluated in data.

Acknowlegdements

First, I would like to thank my supervisor Troels C. Petersen, for his unrelenting motivation and stream of ideas that shaped the project for the better, and for establishing an environment enabling me to work closely with talented individuals.

Next, I have to thank Daniel Nielsen who acted as my unofficial co-supervisor. His brilliant ideas and willingness to help, no matter the problem, made him an absolutely irreplaceable office mate.

Finally, I would be remiss if I did not place a massive thanks to the office crew: Daniel Nielsen, Lukas Ehrke, Christian Michelsen and Frederik Faye.

Introduction

This thesis builds on two main motivations, the first of which is improving photon identification in the ATLAS detector and the second is testing the waters for data training on ATLAS data. The work presented will show the results of Monte Carlo trained boosted decision trees and feed forward neural networks, for photon identification. Evaluated both in Monte Carlo and in data. It will show a data-driven framework for producing a measure of isolation decorrelated from identification and an application of selecting cleaner labels for training in data. Resulting in data-driven electron and photon identification classifiers. The performance of all classifiers have been compared to the ATLAS Loose and Tight working points, on truth labels and the Higgs peak in Monte Carlo and on $Z \rightarrow ee$ and $Z \rightarrow ll\gamma$ datasets in data. The structure of the thesis will be as follows: In chapter 1, the standard model and relevant particle physics concept and terminology will be introduced. Chapter 2 will in the same essence introduce: the experimental setup, the ATLAS detector, its subdetectors, and the current method for which ATLAS reconstructs and identifies electrons and photons. Chapter 3 will describe the different datasets used for training and evaluation and how they were selected. Chapter 4 will shortly introduce the machine learning models and the frameworks used to assist the process and training, optimization and evaluation. Chapter 5 will describe all the results and the process with which they were obtained. This chapter aims to walk through the work done from beginning to end, taking a step back to describe the important details, when necessary. Before chapter 6 collects and summarizes the results, while looking beyond the scope of the project at the next steps, should they be short or long.

Theory

The Standard Model of Particle Physics

The SM is the current best explanation for the nature of particle physics. It predicts the existence of the elementary particles found in figure 1 and experiments have proven the existence of all of them. The SM encompasses the following quantum field theories (QFTs): Electroweak theory (EW) and Quantum Chromodynamics (QCD)

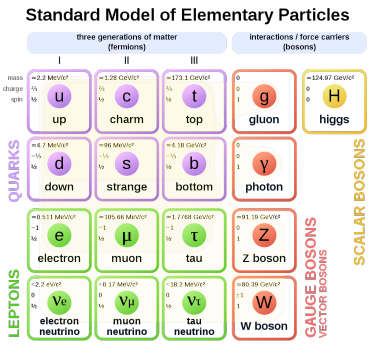


Figure 1: The particles in the Standard Model of physics.

Interactions

Along with the particles, the standard model predicts a list of interactions and self-couplings found in figure 2

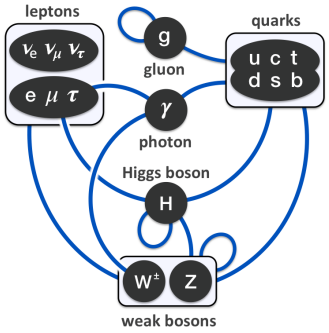


Figure 2: The interactions between particles predicted by the Standard Model.

Units

It is worth taking a step back before heading into the physics more closely related to the process of colliding and measuring particles in modern colliders to describe some useful concepts related to units and constants in High Energy Particle Physics (HEP).

Firstly, for convenience we use *natural units*, wherein $\hbar = c = 1$. Meaning energies, masses and momenta are measured in energy, namely Electron volts (eV) ($1eV = 1.602 \times 10^{-19} J$), whereas lengths and times are measured in eV^{-1}.

Collider Physics

The standard coordinate system of a collider has the z-axis along the beampipe, with the positive x-axis pointing towards the center of the Large Hadron Collider (LHC) and the positive y-axis pointing upwards. The azimuthal angle ϕ is measured around the z-axis and the polar angle θ is the angle with respect to the z-axis. The rapidity, defined as $y = \frac{1}{2} ln[\frac{(E+p_z)}{(E-p_z)}]$, which for massless particles reduces to $\eta = -ln(tan(\theta/2))$, is almost exclusively used instead of θ.

A measure of the interaction rate of colliding particles is the cross section, σ, acquired by applying a solid angle integral to the following formula:

$$\frac{d\sigma}{d\Omega} = \frac{1}{I_b \rho dx} \frac{dL\sigma}{d\Omega} \tag{1}$$

where I_b is the current of beam particles, ρ is the density of the target, dx is the target thickness and L is the luminosity. It describes the effective cross section of interaction and is measured in barns ($1barn = 10^{-24}cm^2$), it has the dimension of area.

$$\frac{N_{events}}{s} = L\sigma \quad , \quad L = \frac{n_{bunches}N_1N_2f_{rev}}{A} \tag{2}$$

Where $N_{1(2)}$ is the number of events in the corresponding bunch, A is the overlapping area of the two bunches colliding head on, and f_{rev} is the frequency of revolution. Increasing luminosity is done by increasing the number or the content of the bunches, or by minimizing the area of each bunch. Ultimately, the luminosity of your collider is the first measurement of your ability to probe physics, given that it directly correlates to the amount of statistics available to your experiment. The given conditions of your experiment changes over time, which leads to two different definitions of luminosity: an instantaneous luminosity calculated as equation 2 dictates with the current conditions and an integrated luminosity, which is calculated by doing a time integral of the instantaneous luminosity.

The integrated luminosity of the ATLAS experiment over the data taking period spanning the 3 years of 2016 to 2018[1] is shown in figure 3.

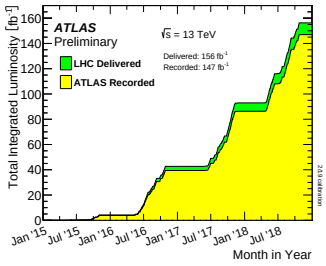


Figure 3: The accumilative integrated luminosity of the ATLAS detector during Run2. Figure taken from **[noauthor_luminositypublicresultsrun2_nodate]**

[1] Known as Run2

Charged and Neutral Interactions with matter

Charged and neutral particles traversing matter behaves very differently and are best described separately. The goal is to understand how particles can interact with their surroundings as they traverse space. This is valuable information when one tries to understand, not only the behavior of particles in the ATLAS detector, but also the design of the detector itself. The detector will be described in detail later. The main focus are on electrons and photons but where applicable other particles will also be mentioned.

Charged Particles interact with matter by losing kinematic energy through three processes: ionization, bremsstrahlung, and multiple scattering. The latter being elastic scattering with nuclei, it only causes a significant energy loss for electrons, however, it does cause a change of direction for particles of all masses. This contributes to momentum measurement uncertainties. Energy loss to bremsstrahlung is severely suppressed for projectiles much heavier than the electron and its effect will be described later. This leaves ionization as the only viable method for charged particles of significant mass to lose energy when traversing matter. The quantum mechanical expression of ionization is the Bethe-Bloch formula, which dictates that energy loss due to ionization depends heavily on the particle β-factor[2] and charge. The main other dependence is the stopping power[3] of the material it passes through. The dependence on β causes the particle to lose energy very fast when it drops below 1. The energy deposit of

[2] v/c

[3] Depends itself on atomic number and weight, mass density, electron configuration and such.

the the particle will therefore be very localized space, in a so-called Bragg peak. Given the high energy of the particles in the ATLAS detector, ionization is only the absolute final contribution with which they are stopped. It is, however, used avidly for determining their trajectory in the detector.

The electron is by far the lightest electrically charged particle, with the second in line being the muon of 200 times the electron mass. This makes electrons a special case, as they are open to interact in multiple ways with matter. They ionize very similarly as described above, however, their low mass forces them to be deflected more when traversing material and quantum mechanics does not allow for us to distinguish traversing electrons from the ones liberated. This complicates the ionization calculation but the dependencies remain the same. However, for an electron commonly found in the ATLAS detector, the dominant method for which they lose energy is through bremsstrahlung, the radiation of a photon caused by interactions with the atomic nuclei. The Coulomb forces exerted on the electron cause a sudden acceleration, which results in a photon being radiated. The likelihood of this happening mainly depends on the amount of protons in the nuclei (charge) and the degree to which the electron configuration shrouds[4] the nuclei. From here a measure of radiation length, X_0, is defined, which is a characteristic of material. It describes the average length traveled by an electron before it loses all but $1/E$ of its energy, called the mean free path. Muons travel almost exclusively as a Minimally Ionizing Particle (MIP), meaning they leave only small deposits of energy as they traverse material and is therefore extremely hard to stop. However, given they are next in line to lose energy through bremsstrahlung, muons will also radiate photons above an energy threshold. For lead this threshold is $\sim 300GeV$.

[4] Usually referred to as screening

Finally, charged particles will radiate photons as they move through inhomogeneous materials, specifically as they transition through material boundaries. This is called transition radiation and is highly depending on the particles' Lorentz factor $\gamma = E/m$. This effect is once again almost exclusively viable for electrons.

From this it should be very clear that, especially, electrons produce photons as they traverse matter. Knowing how these photons will interact is therefore vital to understanding the behavior of not only the photons produced in hard-scatter events, but also the photons produced through secondary processes all throughout the detector.

Neutral Particles The photon interacts with matter very differently depending on the energy of the photon. The processes relevant to the ATLAS environment are Compton scattering and pair production. Compton scattering is dominant at the order of $10MeV$ and is elastic scattering with electrons. Pair production is by far the dominant process in the ATLAS environment and is quantum mechanically caused by the photon fluctuating to a virtual e^+e^- pair. The pair is brought on shell through exchange of a photon with a nucleus. Pair produc-

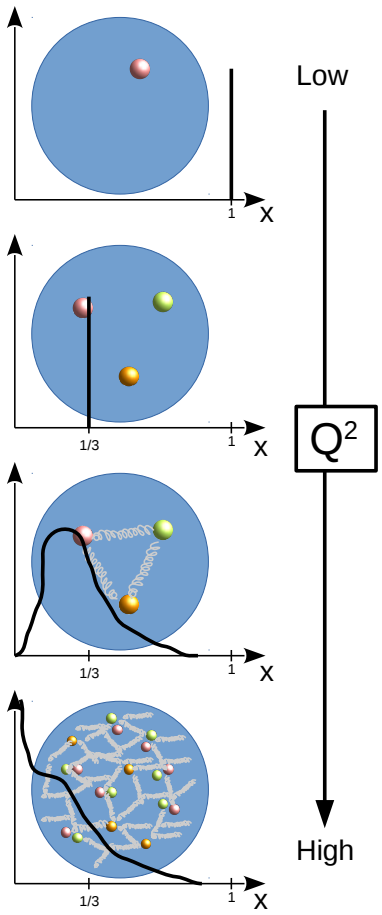


Figure 4: Visualization of the proton contents and corresponding PDF as a function of Q^2. The bjorken x represents the fraction of longitudinal momentum carried by a given parton

tion is very related to bremsstrahlung and also depends heavily on the screening effect of the electron surrounding the nucleus. In fact, the mean free path of photons are $\frac{9}{7}X_0$, slightly longer than the electron mean free path. We now essentially have a recipe for how a high energy photon and electron will dispose of the vast majority of their energy. The electron will eventually radiate a photon, and the photon will eventually decay into an electron-positron pair. From here this process will simply repeat itself with the energy contained falling exponentially. This cascade is called an electromagnetic shower and the only difference between the shower of an electron and a photon is the slightly longer *average* depth before the shower starts.

All the covered interactions were electromagnetic of nature and as explained many of these processes are highly suppressed for charged particles with a high mass. The most commonly produced particles in modern collider experiments are pions (π^0, π^+, π^-). These, along with other mesons and hadrons, interact mainly through interactions with nuclei in a short range inelastic strong force interaction. The mean free path is referred to as Nuclear Interaction lengths (NIs) and is much longer than X_0 for a given material. The shower produced after the first interaction is of much more complicated nature than the showers of electrons and photons. However, they often contain an electromagnetic component due to production of π^0's that quickly decay into photons. The hadronic component of the shower is more chaotic, mainly due to the many possible strong interactions. For reference the material characteristics of lead yields $X_0 = 0.5612cm$ and $NI = 17.59cm$[**noauthor_notitle_nodate**].

Probability Distribution Functions (PDFs)

The proton is not an elementary particle and is canonically made up of three valence quarks, however, the wonderful chaotic nature of QCD dictates that reality is much more complicated than this. The proton also contain sea quarks and gluons created through QCD processes. These along with the valence quarks are referred to as partons. The structure of the photon greatly changes depending on the energy scale, Q^2, as visualized by figure 4. This is an approximation made by defining the proton in an infinite momentum reference frame, which holds well for high scales. The probability of finding a parton carrying a fraction of momentum, x, at a given energy scale, Q^2, is described by PDFs. The energy scale of the LHC is of the order $Q^2 = 10^4 GeV$ and the relevant PDF is found in figure 5, along with one of a much lower energy scale for reference. It is statistically unlikely to produce collisions of partons with a high x and therefore it is clear from figure 5 why the LHC is often referred to as a gluon-gluon collider.

Pileup

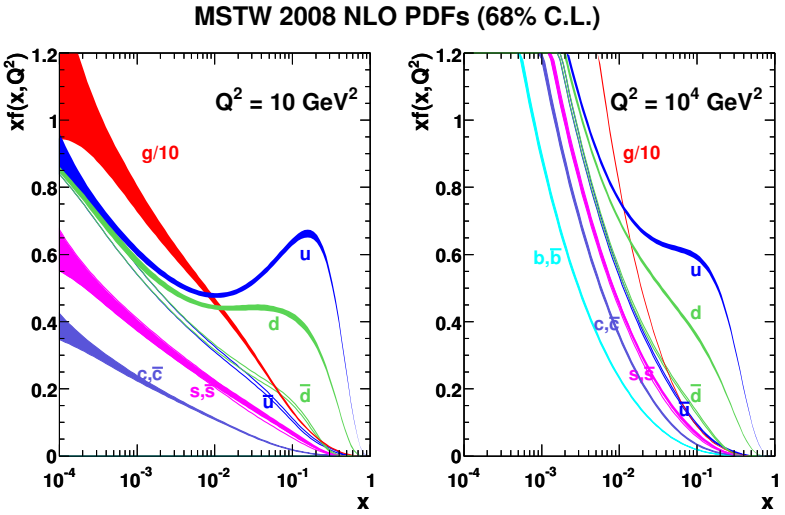


Figure 5: PDFs at two different energy scales Q^2. Figure taken from [**Martin_2009**]

A proton-(anti)protron collider in reality collides a long list of partons, that interact in many different way. The PDFs described in the previous section shows the probability being by far the highest for a gluon-gluon interaction. In reality, due to the bunch structure, not just one interaction happens at each bunch crossing but many. The number of interactions per crossing are recorded across roughly every minute of data taking[5]. The distribution of interactions per crossing across Run2 is shown in figure 6, along with the average interactions per crossing, $\langle\mu\rangle$, which is clearly increasing. It is rare for a given of these interactions to produce a so-called hard interaction, which are head-on interactions with a significant transverse momentum transfer ($pT = psin(\theta)$) in the final state. This type of event is responsible for almost all interesting physics in ATLAS. The opposite is a soft[6] interaction, where no significant transverse momentum transfer found place. These can be long distance elastic scattering between the colliding protons, where the protons simply change direction. They can also be diffractive, which means that either both or one of the protons break up sending a spray of particles going down the beampipe. Finally, they can be non-diffractive, resulting in uniform energy deposits in the calorimeter and no preference in phi. Soft interactions makes up $\sim$ 99.999% of the interactions at the LHC. The large discrepancy between the frequency of hard and soft interactions and the high number of interactions per crossing, makes it impossible for hard interactions to not be accompanied by soft interactions. This is referred to as pileup. Included in this term is the underlying event, which is the partons remaining from the hard interaction. Pileup is then identified as in- or out of time. The former is from other interactions in the same bunch crossing and the latter is from adjacent bunch crossings.

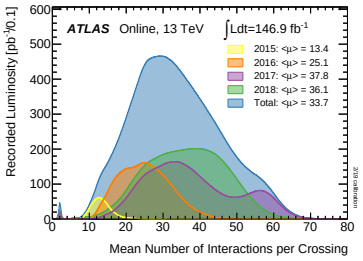


Figure 6: The integrated luminosity recorded at a given mean interactions per crossing in ATLAS for the whole of Run2. Figure taken from **[noauthor_luminositypublicresultsrun2_no**

[5] Referred to as a LumiBloek

[6] This is not an exact definition

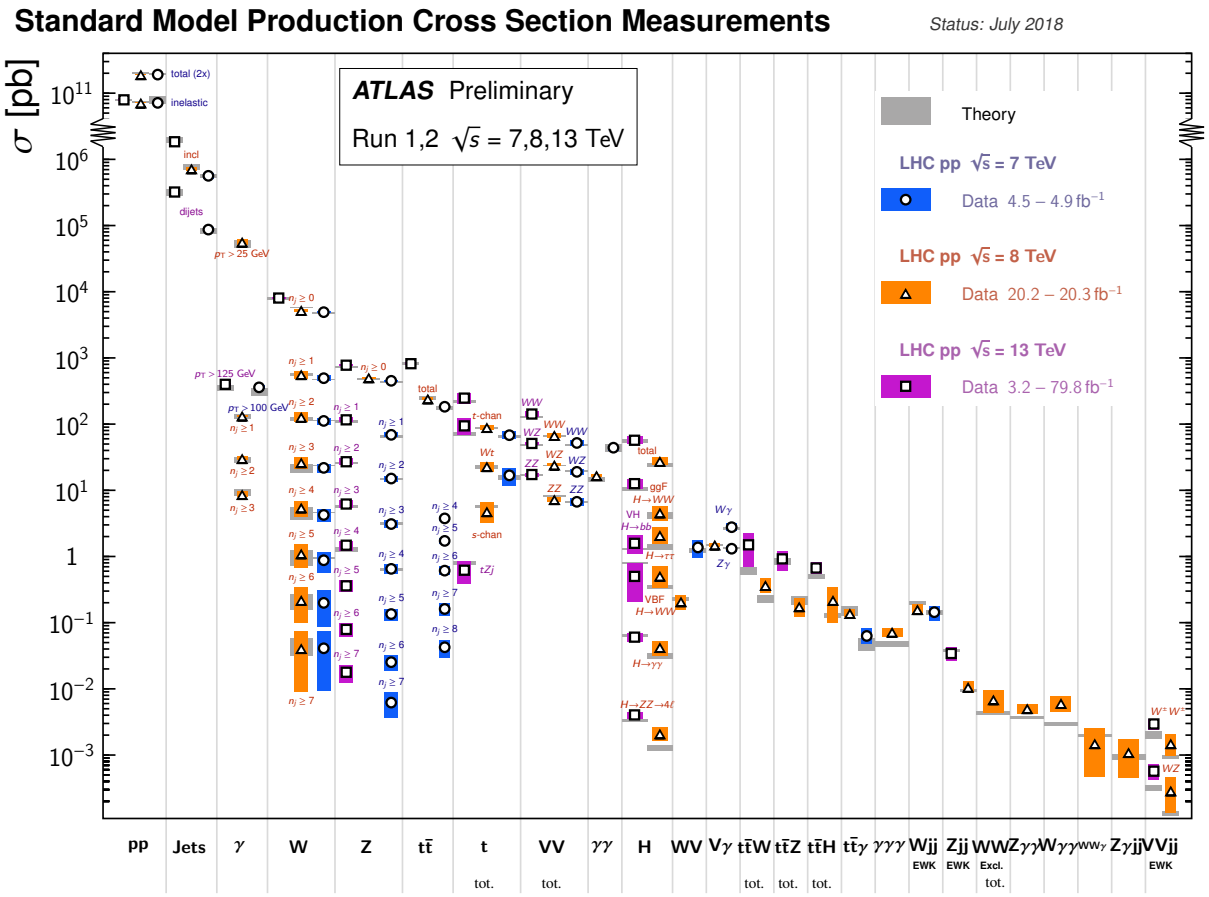


Figure 7: Summary of production cross sections measurements for different SM processes with the uncertainties and integrated luminosity used for the measurement indicated by the markertype and colorbar. Figure taken from [**noauthor_summary_nodate**]

Cross Sections at the Large Hadron Collider

The summary of production cross sections measured at ATLAS can be found in figure 7. One recognizes the cross section from equation 2, showing the simple relationship between cross section, luminosity and how many events one would expect over time. If one inserts the production cross section a given process, one can easily calculate the expected number of events of that process. For example, given an integrated luminosity of $79.8 fb^{-1}$ and a production cross section for the Higgs boson of $\approx 50pb$, one would expect to produce roughly 4 million Higgs bosons. However, for a channel such as $H \to \gamma\gamma$ the production cross section is only $\approx 0.07pb$, given the low branching ratio of the channel[7]. The channel actually only contains in the order of 5000 Higgs bosons.

[7] The probability for Higgs to decay into two photons

$$\sigma = \sum_{a,b=q,g} \int_0^1 dx_a dx_b f_a(x_a, \mu^2) f_b(x_b, \mu^2) \times \frac{1}{2\hat{s}} \int_{cuts} \prod_{i=1,n} \frac{d^3 p_i}{2E_i (2\pi)^3} (2\pi)^4 \delta^{(4)}(p_a + p_b - \sum_i p_i) \bar{\sum} |\bar{M}^{ab\to 1,\ldots,n}|^2 \tag{3}$$

The master formula for production cross sections are the LHC is equation 3. It has a great number of details and will be described in broad strokes, to give an insight into where the numbers from figure 7 come from and how calculations such as the one done above can be so simple. $\hat{s} = (p_a + p_b)^2 = x_a x_b s$, where x is the Bjorken x introduced earlier and s is the collision momentum. $\mu^2 = Q^2$. The cuts represent the reduction of phase space made by applying cuts to for instance E_T and η of the particles. n represents the number of particles in the final state, and $s\bar{u}m|\bar{M}^{ab\to 1,\ldots,n}|^2$ represents the strength of interaction ($M = \langle f|H_I|i\rangle \sim \int e^{i\vec{q}\cdot\vec{r}} V_I(r) d\vec{r}$)[8]. Equation 3 expresses the cross section for two hardons, be they gluons or quarks, to produce a final state of n specific particles. The first term expresses the probability of two partons at the relevant scale, Q^2, to interact, while the second term expresses the production cross section of the specific process. Naturally, the first term depends on the PDFs, while the second term depends on interaction strengths. This is possible due to the perturbative nature of close range QCD interactions, while not being possible for the long distance interactions of diffractive elastic scattering.

[8] The Matrix element from Fermi's golden rule, calculated using Feynmann rules of interaction

The ATLAS experiment

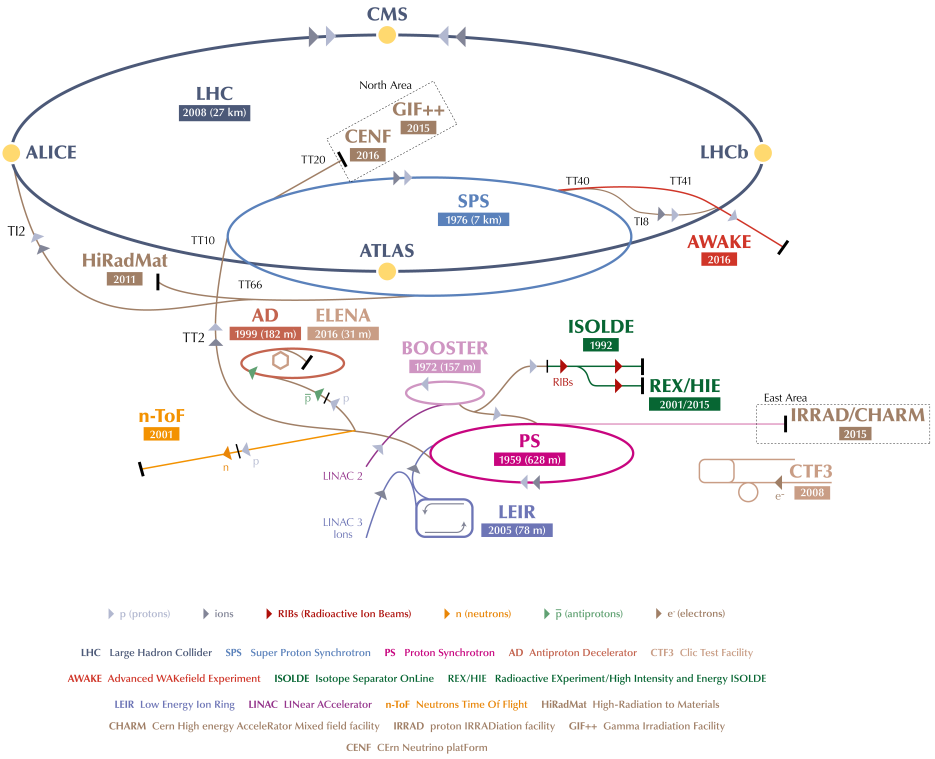


Figure 8: An overview of the CERN accelerator complex in the region around Geneva, Switzerland.

The Large Hadron Collider

Protons from a simple canister of hydrogen part ways with their electron and begin a journey that brings their velocity to 99.999999% the speed of light and energy to 6.5TeV, through the linear accelerator, LINAC2, the circular Proton Synchrotron Booster (BOOSTER), Proton Synchrotron (PS), Super Proton Synchrotron (SPS) after which they are injected into the Large Hardron Collider at the energy of 450 GeV. The final part of the journey includes at least 18 million revolutions around the 27 kilometer ring until it terminates in a collision ($\sqrt{s} = 13TeV$) with another proton in the center of one of the four

major experiments, A Torodial LHC ApparatuS (ATLAS)[**Aad:2008zzm**], Compact Muon Solenoid (CMS)[**Chatrchyan:2008aa**], A Large Ion Collision Experiment (ALICE)[**Aamodt:2008zz**], and Large Hadron Collider beauty (LHC-b)[**Alves:2008zz**]. The first of which is the focal point of this thesis.

The LHC[**Evans:2008zzb**] is an incredibly complicated machinery and the next section will therefore only briefly describe the main components that make this whole thing possible, including the three concepts; acceleration, bending and focusing.

The protons are structured in bunches containing around 10^{11} protons (10^9 anti-protons) and at full capacity the LHC contains maximally 2556 bunches with a $25ns$ gap between bunches. The acceleration is generated by radio frequency (RF) cavities. An oscillating electric field, which in turn attracts protons heading towards the cavity and repels protons moving away. This both leads to acceleration and help maintain the bunch structure of the beam, since the acceleration is based on how well in tune you are with the cavities. This method of acceleration is used in all steps from the LINAC2 to the LHC. In a linear accelerator the RF cavities come in a line (with increasing increments to accommodate the acceleration), resulting in each cavity only being used once. In circular colliders acceleration only happens at certain points in the accelerator and they are reused many, many times. This is made possible by introduction of a magnetic field, which bends the trajectory of the particles in an arc. As acceleration happens this magnetic field is increased in strength. In the case of the LHC an 8 Telsa field supplied by super conducting magnets operating at 1.9 K is needed.

The beam is kept focused by use of FODO cells (Focus-drift-DeFocus-drift), which is achieved by using a quadrupole doublet, essentially acting as a focusing lens followed by a defocussing lens with a careful chosen drift-space in between. The quadrupole magnets work by producing a magnetic field that increases moving radially out from the center bending the trajectory inwards for focusing and outwards for defocussing. The first quadrupole magnet bends the beam inwards and the second magnet bends the beam outwards, the net result is a trajectory parallel to the direction of movement and more collimated than before. This works essentially as a convex lens followed by a concave lens but by manipulating charged particles instead of light.

These three mechanics provide the foundation needed for high energy, luminosity, and stable operation. Together with an endless list of mechanics that are equally vital in making this whole machinery work, e.g. injection, cryogenics, dumping of beam and many more. The LHC has continued to operate beyond expectation providing the connected experiments with unprecedented amounts of statistics and has been pushing the boundaries of particle physics ever since it begun operation in 2008.

The ATLAS experiment

One of these experiments, ATLAS, is the foundation upon which this thesis was built. The knowledge of how the detector works is therefore of utmost importance and while it is impossible for me to cover all the details of the detector. Providing a general overview with important details will be the goal of this section and the information outlined is paraphrased from the ATLAS design paper[**Aad:2008zzm**]. The collisions described in chapter happen in the center of the detector. Along with CMS, ATLAS is designed as a general-purpose particle physics experiment. This means it has to capable of precisely measuring and reconstructing all final state particles produced by collisions in the kinematic range allowed by the LHC. This includes known, but certainly also unknown, types of events. One of the prime goals of the experiment was to find the Higgs boson, a long missing piece of the standard model. Today, long after the discovery of the Higgs boson in 2012, the experiment and its massive collaboration continues to use the incredible statistics gathered to probe and set limits on increasingly rare processes and couplings.

The subdetectors of ATLAS

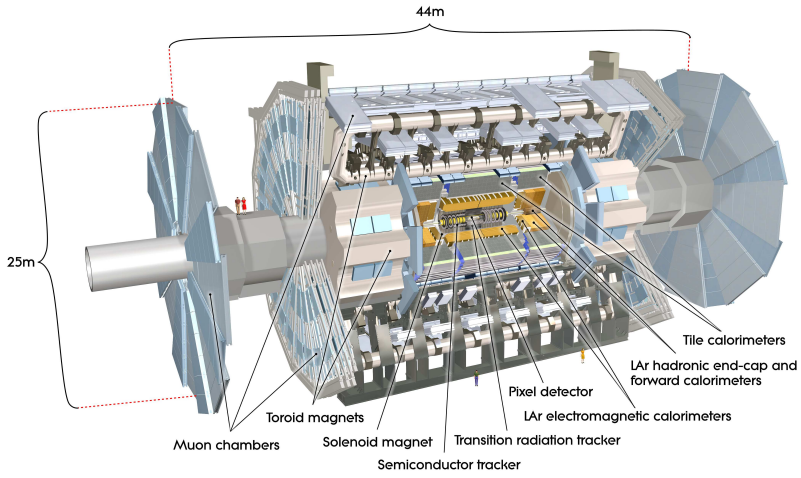


Figure 9: A cut out view of the ATLAS detector.

The ATLAS detector, shown in figure 9, is a very complicated machine, built of many smaller, equally complicated machines, referred to as subdetectors. The objects of interest to the work done in this thesis mostly relate to the active detector material. However, knowledge of the many nuances and quirks of both the active and inactive material has shown to be important. Here activity refers material actively collecting measurements, as opposed to power cables, cooling, and so forth. Everything inside the detector has a purpose. The individual subdetectors all serve a carefully designed purpose in order for the complete detector to acquire all of the needed versatility for the task at hand. The major subdetectors are the ID, the ECAL,

the Hadronic Calorimeter (HCAL) and the Muon Systems (MS), all of which themselves have multiple subdetectors covering different regions of $|\eta|$.

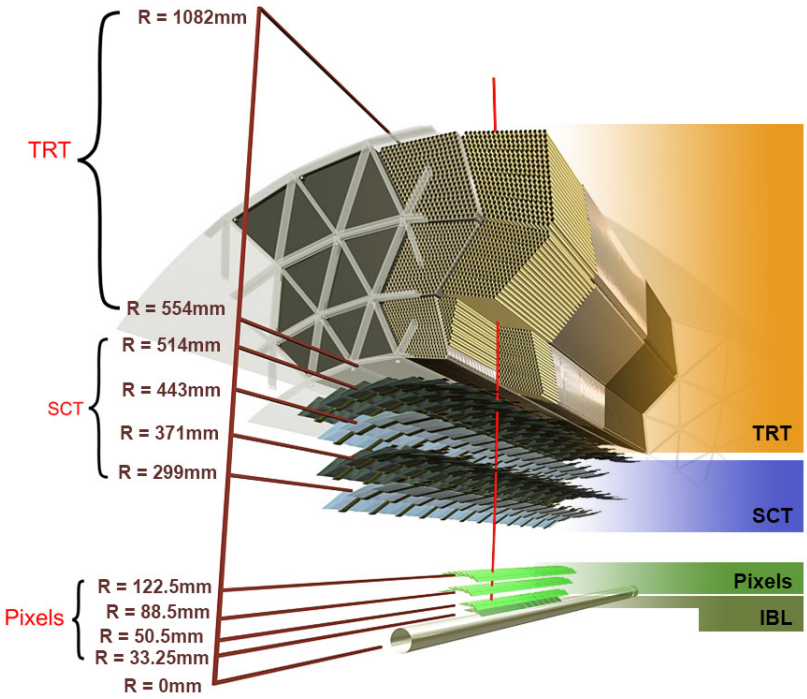


Figure 10: A computer generated image of the Inner Detector, the subdetector of the ATLAS detector responsible for tracking of charged particles. The subdetectors of the ID are the Insertable B-Layer (IBL), silicon pixel layers (Pixels), the SCT and the TRT. Figure taken from [**potamianos2016upgraded**]

The Inner Detector is responsible for the tracking of charged particles. It is positioned closest to the beampipe and contains multiple subdetectors moving out radially: the silicon pixel layers, the silicon-microstrip sensors (both covering $|\eta| < 2.5$), and the transition radiation tracker (covering $|\eta| < 2.0$). Spanning the detector is a 2 Telsa solenoidal magnetic field, with the purpose of bending the trajectory of charged particles. This allows for momentum measurements, determination of their charge and both primary and secondary vertex measurements. The silicon pixel layers contain 1744 identical pixel sensors, each containing 47232 pixels of nomial size $50 \times 400 \mu m^2$. However, for spatial reasons the total number of readout channels for each pixel is 46080, totaling almost 80.4 million readout channels. In long shutdown 1 (LS1), in-between Run1 and Run2, the pixel detector received an upgrade in the form of an additional layer, the Insertable B-layer, closest to the beampipe with a resolution of $50 \times 250 \mu m^2$, adding a total of 12 million pixels. The original ID sensors are spread across three barrel layers and three end-cap disk layers (mirrored in η). The SCT provides at least four layers of coverage across the whole fiducial area covered by the ID. This is achieved by 4 barrel layers and 9 end-cap layers. It contains 2012 modules of $12cm$ length in the barrel and 1976 modules in the end-cap of have varying lengths to give consistent spatial resolution. Totaling 4088 modules, each containing 2×768 active strips, totaling almost 6.3

million readout channels. The ID subdetectors all work based on semiconductors. They consist of a p-n junction with reserve bias applied, creating a depletion zone containing no free carriers. A charged particle traversing this depletion zone will create free carriers by ionization. They will move along the electric field and create a current in the shape of a pulse. The massive amount of channels looking for these pulses gives the ID the small spatial resolution needed to identify and measure the tracks of the many charged particles that traverse the detector.

The TRT uses the concept of transition radiation, described in chapter . This detector is specifically added to improve electron identification. In the barrel region it contains up to 73 layers of straws interleaved with fibers, while in the end-cap it contains 160 straw planes interleaved with foils. The fibers and foil are transition radiation material and the design of the detector attempts to force transition radiation by creating many transitions between materials. As vaguely noticeable in figure 10, the TRT in the barrel is segmented into three rings of 32 modules each. The straws are staggered and have a mean spacing of $\sim 7mm$. The end-cap is mirrored in η and consists of two independent sets of wheels, with the inner most wheel containing 12 wheels and the outer one containing 8. All wheels contain 8 layers of straws separated by foil. The straws themselves have a wire drawn in the middle across their full length. An electrical potential is then applied across the straws and the wire along their center. The entire TRT is filled with a careful chosen mixture of gas, a noble gas for maximum amplification, Xenon ($\sim 70\%$)[9], a quencher gas to lower ultra-violet radiation from exited atoms and ions, CO_2 ($\sim 27\%$), and a gas, O_2 ($\sim 3\%$), to catch electrons moving unintentionally long distances. A charged particle traversing the straws will ionize the Xenon, creating ion pairs on the order of 300 per cm. This creates an avalanche of drift electrons towards the center wire, ultimately creating an electrical current in the shape of a sudden pulse with a long tail. The signal has a pulse time of $\sim 20ns$ with an additional $\sim 8ns$ rise time, which provides the necessary conditions for operation in the LHC. A charged particle will typically leave 36 of such pulses (hits) in the TRT. The hits provided by the TRT has lower spatial resolution than the hits in silicon detectors. However, the sheer number of hits assists track reconstruction and electron identification. In total, the TRT has approximately 351000 readout channels. A 'hit' in the ID constitutes a signal above a carefully selected threshold and will be used exclusively when describing interaction with the ID going forward.

An overview of the ATLAS calorimetry and its subdetectors coverage in η as well as granularity given in $\Delta\eta \times \Delta\phi$ ($\Delta x \times \Delta y$ the FCal) can be found in table 1.

[9] Due to leakage some modules now contain 70% Argon instead[**collaboration2017performance**]

EM Calorimeter Barrel			
	$\lvert\eta\rvert$ Coverage	Granularity $\Delta\eta \times \Delta\phi$ (versus $\lvert\eta\rvert$)	
Presampler	$\lvert\eta\rvert < 1.52$	0.025×0.1	
Calorimeter 1st layer	$\lvert\eta\rvert < 1.475$	$0.025/8 \times 0.1$	$\lvert\eta\rvert < 1.40$
		0.025×0.025	$1.40 < \lvert\eta\rvert < 1.475$
2nd layer	$\lvert\eta\rvert < 1.475$	0.025×0.025	$\lvert\eta\rvert < 1.40$
		0.075×0.025	$1.40 < \lvert\eta\rvert < 1.475$
3rd layer	$\lvert\eta\rvert < 1.35$	0.050×0.025	
EM Calorimeter End-cap			
Presampler	$1.5 < \lvert\eta\rvert < 1.8$	0.025×0.1	
Calorimeter 1st layer	$1.375 < \lvert\eta\rvert < 3.2$	0.050×0.1	$1.375 < \lvert\eta\rvert < 1.425$
		0.025×0.1	$1.425 < \lvert\eta\rvert < 1.5$
		$0.025/8 \times 0.1$	$1.5 < \lvert\eta\rvert < 1.8$
		$0.025/6 \times 0.1$	$1.8 < \lvert\eta\rvert < 2.0$
		$0.025/4 \times 0.1$	$2.0 < \lvert\eta\rvert < 2.4$
		0.025×0.1	$2.4 < \lvert\eta\rvert < 2.5$
		0.1×0.1	$2.5 < \lvert\eta\rvert < 3.2$
2nd layer	$1.375 < \lvert\eta\rvert < 3.2$	0.050×0.025	$1.375 < \lvert\eta\rvert < 1.425$
		0.025×0.025	$1.425 < \lvert\eta\rvert < 2.5$
		0.1×0.1	$2.5 < \lvert\eta\rvert < 3.2$
3rd layer	$1.5 < \lvert\eta\rvert < 2.5$	0.050×0.025	
Barrel (Extended barrel) Tile Calorimeter			
1st layer	$\lvert\eta\rvert < 1.0$ ($1.1 < \lvert\eta\rvert < 1.7$)	0.1×0.1	
2nd layer	$\lvert\eta\rvert < 1.0$ ($1.0 < \lvert\eta\rvert < 1.6$)	0.1×0.1	
3rd layer	$\lvert\eta\rvert < 1.0$ ($0.8 < \lvert\eta\rvert < 1.5$)	0.2×0.1	
Hadronic End-cap Calorimeter			
Presampler	$1.5 < \lvert\eta\rvert < 3.2$	0.1×0.1	$\lvert\eta\rvert < 2.5$
		0.2×0.2	$\lvert\eta\rvert > 2.5$
1st layer	$1.5 < \lvert\eta\rvert < 3.1$	0.1×0.1	$\lvert\eta\rvert < 2.5$
		0.2×0.2	$\lvert\eta\rvert > 2.5$
2nd layer	$1.6 < \lvert\eta\rvert < 3.2$	0.1×0.1	$\lvert\eta\rvert < 2.5$
		0.2×0.2	$\lvert\eta\rvert > 2.5$
3rd layer	$1.7 < \lvert\eta\rvert < 3.3$	0.1×0.1	$\lvert\eta\rvert < 2.5$
		0.2×0.2	$\lvert\eta\rvert > 2.5$
Forward Calorimeter			
1st layer	$3.0 < \lvert\eta\rvert < 5.0$	3.0×2.6	$3.0 < \lvert\eta\rvert < 4.3$
		~ 4 times finer	$3.10 < \lvert\eta\rvert < 3.15$
			$4.3 < \lvert\eta\rvert < 4.83$
2nd layer	$3.1 < \lvert\eta\rvert < 5.0$	3.3×4.2	$3.24 < \lvert\eta\rvert < 4.50$
		~ 4 times finer	$3.20 < \lvert\eta\rvert < 3.24$
			$4.50 < \lvert\eta\rvert < 4.81$
3rd layer	$3.2 < \lvert\eta\rvert < 5.0$	5.4×4.7	$3.32 < \lvert\eta\rvert < 4.60$
		~ 4 times finer	$3.29 < \lvert\eta\rvert < 3.32$
			$4.6 < \lvert\eta\rvert < 4.75$

Table 1: The full overview of the ATLAS calorimetry with the coverage of each layer and their granularity in $\Delta\eta \times \Delta\phi$ ($\Delta x \times \Delta y$ for the FCAL). Table gathered from [**Aad:2008zzm**]

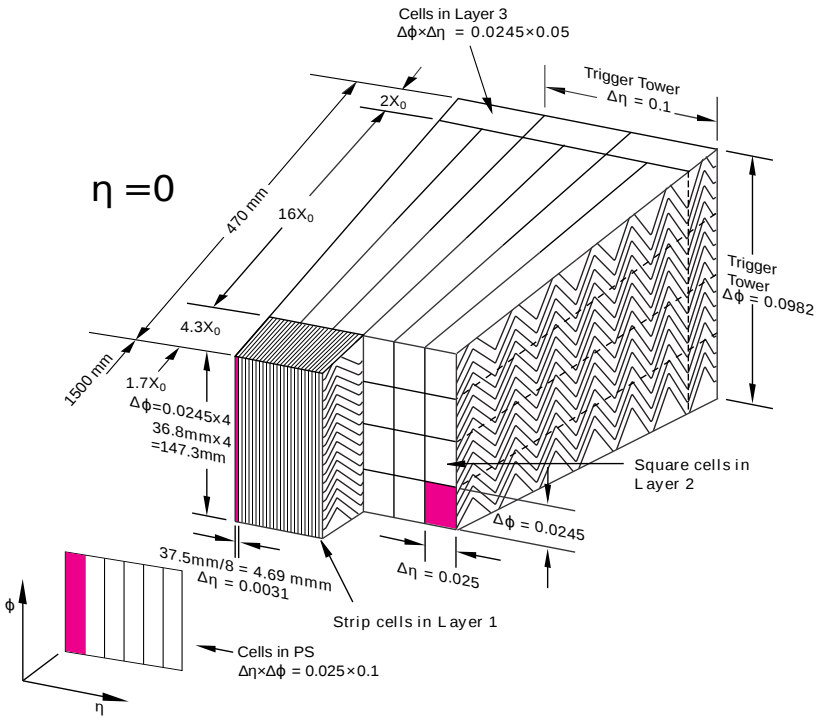


Figure 11: ATLAS Electromagnetic Calorimeter (barrel) module. The layered accordion structure as well as the $\Delta\eta \times \Delta\phi$ granularity can be clearly seen. Also indicated is the respective physical and radiation depth of the module. Figure taken from **[Aaboud_2019]**

The Electro-Magnetic Calorimeter is separated in two distinctly different calorimeters differentiated by their position in η. The barrel ECAL and the end-cap ECAL. For both calorimeters the active detector material is Liquid Argon (LAr) and the absorber material is lead. The barrel ECAL consists of two half-barrels connected at $z = 0$, one covering the region $z > 0(-1.475 < \eta < 0)$ and the other $z < 0(0 < \eta < 1.475)$. A half-barrel consists of 1024 accordion-shaped absorbers and is divided into 16 modules covering $\Delta\phi = 22.5°$ each. One such module is shown in figure 11. The module shown is at $\eta = 0$ has a depth of 22.3 radiation lengths (X_0, introduced in chapter). An overview of the radiation lengths, X_0, of whole the ECAL is shown in figure 12. In front of the active EM calorimeter is an extra layer called the presampler, as shown in 11. It provides full coverage of the barrel ECAL with a resolution ($\Delta\eta \times \Delta\phi$) of 1.52×0.2.

On each side of the barrel ECAL two co-axial wheels, which makes up the end-cap calorimeters (EMEC). They covers the region $1.375 < |\eta| < 3.2$, with a boundary between the two co-axial wheels at $|\eta| = 2.5$. The inner and outer wheel contains 256 and 768 absorbers, respectively. The depth increases from 24 X_0 to 38 X_0 in the region $1.475 < |\eta| < 2.5$ (outer wheel) and from 26 X_0 to 36 X_0 in $1.475 < |\eta| < 2.5$ (inner wheel). Once again, a presampler is in front of the active EMEC calorimeter in the limited region of $1.5 < |\eta| < 1.8$.

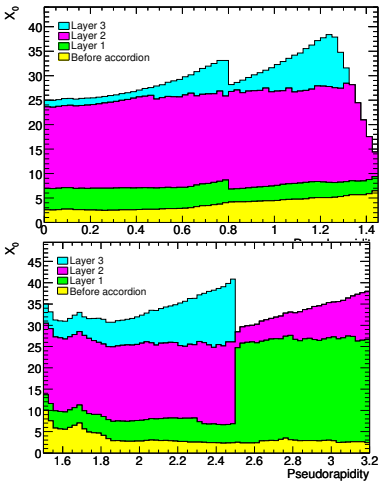


Figure 12: The depth of the ECAL barrel (top) and end-cap (bottom) in radiation length, X_0. Figures taken from **[Aad:2008zzm]**

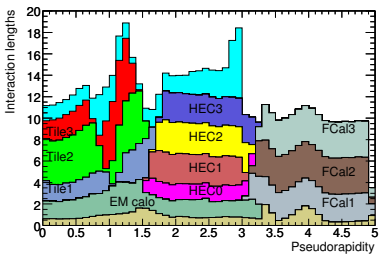


Figure 13: The depth of the Hadronic Calorimetry (and FCAL) in Nuclear Interactions lengths (NI).

The Hadronic Calorimeters Directly behind the ECAL and EMEC calorimeter comes a hadronic calorimeter, namely the tile calorimeter and the hadronic end-cap calorimeter[10]. The former is a scintillator with steel absorbers, while the latter is a LAr calorimeter with copper absorbers. They are both sampling calorimeters, meaning they alternate absorber and active material. For an overview of the depth of the hadronic calorimetry, see figure 13.

[10] It shares cooling system with the EMEC

The tile calorimeter is segmented in two separate detectors of the same type: the barrel tile calorimeter, covering the region $|\eta| < 0.8$, and the extended tile calorimeter, covering the region $0.8 < |\eta| < 1.7$. Given that the detector starts quite far away from origo, changes in the polar angle lead to quite large changes in the physical distance, which indirectly limits the resolution. The granularity is 0.1×0.1 for the first and second layers and 0.2×0.1 for the 3rd layer, applicable for both the barrel and extended barrel subdetectors.

The hadronic end-cap calorimeter (HEC) has a presampler and 3 layers. It covers the region $1.5 < |\eta| < 3.2$ and has a granularity of 0.1×0.1 up until $|\eta| < 2.5$ and 0.2×0.2 for the rest. The elec-

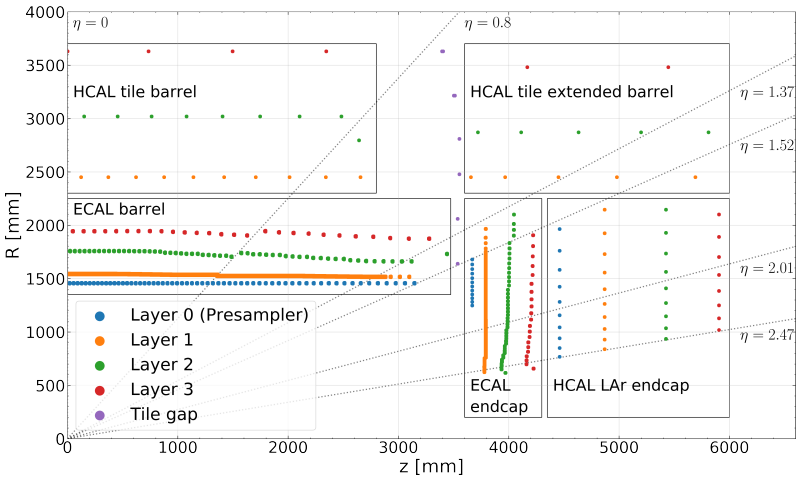


Figure 14: The center of calorimeter cells, labeled according to layer, for the ECAL and HCAL in spacial coordinates. The lines drawn represent the trajectory of a particle of the listed η. The spacial centers of the cells are aquired from the AODs. Figure taken from**[ehrke_machine_2019]**

tromagnetic and hadronic calorimeters are nicely visualized in figure 14, where the center of the cells are plotted along with example tracks of different η values. This puts into context the just described calorimeter positions and granularity.

Forward Calorimeters The high polar angles is an often overlooked part of the detector. The forward region ($3.1 < |\eta| < 4.9$) contains the forward LAr calorimeter (FCal), which consists of 3 layers, one specialized in electromagnetic showers with steel absorbers (FCal1), and two specialized in hadronic showers with tungsten absorbers (FCal2 and FCal3). The depth of the layers are shown in figure 13. This subdetector is exposed to a high flux of particles at a much higher level than the rest of the detector, limiting the design possibilities leading to a rough granularity (1). This, along with a complete lack of tracking, leads to this detector being overlooked in almost

any physics analysis that doesn't specifically deal with this region of phase space.

The task of the calorimetry is to measure the energy of particles. This is accomplished essentially by stopping the particles and measuring the energy deposited as they stop. As described in chapter , the mean free path for photons and electron vary greatly from the mean free path of mesons and hadrons. This allows the ATLAS detector to carefully manage the amount of material present in order to stop electrons and photons in the ECAL. The energy is measured by use of scintillators. Scintillating material dissipates absorbed energy through emission of light and the material is itself transparent to this light. The light is then guided to a photomultiplier tube (PMT), which outputs an electric signal and after careful calibration one can measure the energy deposited.

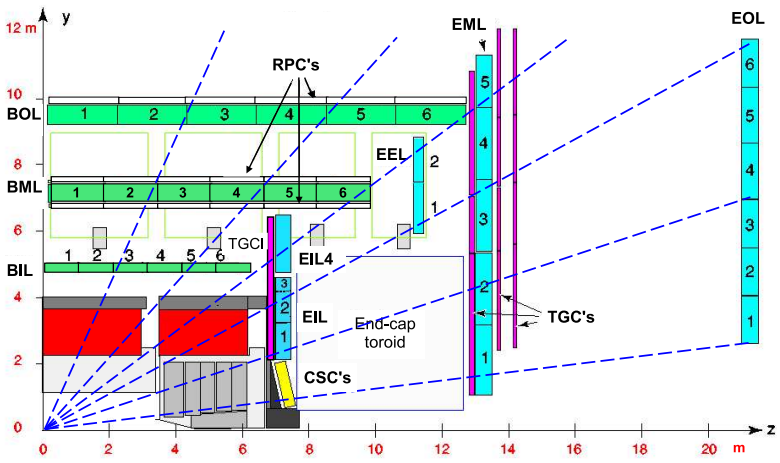


Figure 15: The Muon Systems in spatial coordinates, with the Monitored Drift Tubes (MDT) in green and cyan, the Cathode Strip Chambers (CSC) in yellow, the Resistive Plate Chambers (RPC) in white (attacked to BML and BOL), and the Thin Gap Chambers (TGC) in purple. Infinite momentum muons will follow the blue dashes lines (In reality they bend slightly).

The Muon Systems consists of a muon spectrometer shown in figure 15. Note that the grey and red boxes correspond to the boxes in figure 14. It is designed to identify and measure tracks of muons, who travel through the detector as a MIP and will, therefore, rarely start showers in the calorimeters. For this reason the MS contains four tracking detectors: the Monitored Drift Tubes (MDT), the Cathode Strip Chambers (CSC), the Resistive Plate Chambers (RPC), and Thin Gap Champbers (TGC). The whole system covers a range up to $|\eta| \approx 2.7$ and each detector has a specific purpose. The MDT provides precision tracking inside the whole range, the CSC provides precision tracking in the range $2.0 < |\eta| < 2.7$, the RPC ($|\eta| < 1.05$) and the TGC ($1.05 < |\eta| < 2.7$ (2.4 for triggering)) provides triggering and second coordinate (Non-bending coord.) measurements. Three large air-core toroid magnets, one in each end-cap and one across the barrel, producing a toroidal magnetic field of approximately $0.5T$ and $1T$ in the barrel and end-cap regions, respectively. The muons are measured both as they bend in the inner detector and once more as they bend in the opposite direction in the MS.

This allows precise measurements of the muon kinematics without stopping them.

The Crack In the transition region between barrel and the end-cap is a region of high material density due to cabling and services for the inner detector and barrel LAr calorimeter. It is often referred to as the crack and is in the range $1.37 < |\eta| < 1.52$. In figure 14 this range is shown by η lines. This clearly visualizes the degradation of measurement quality in this specific region. To minimize this degradation several scintillators are installed, they are referred to in as "Tile gap" in figure 14.

Particles in the ATLAS Detector

As described in the previous section, each subdetector in ATLAS serves an individual specifically designed purpose and the goal is to serve as a general-purpose particle physics detector. This is achieved by being able to identify and measure kinematics of every single particle that makes it into the detector. This list includes: electrons, muons, taus, photons, neutrinos, and a range of mesons and hadrons usually collected in hadronization jets.

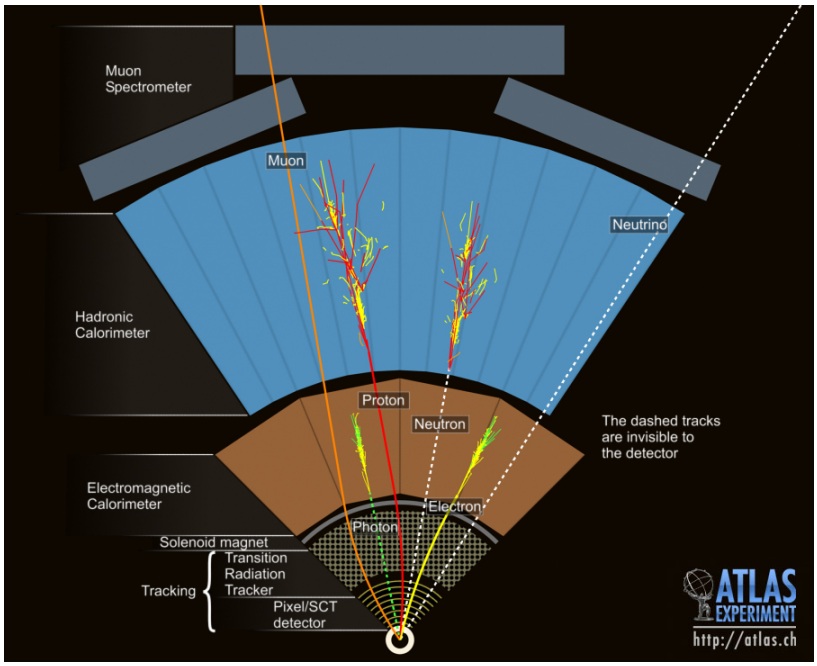


Figure 16: Average trajectories and fates of the common final state particles in the ATLAS detector. Figure taken from **[Pequenao]**

In figure 16 ideal examples of common final state particles are shown. The goal of this section is to understand these ideal cases, since they often lay the foundation of how the detector is designed. Once again, understanding some of all the nuances and caveats that make the world much more interesting than the intentionally simple figure 16 is equally important.

Electrons are the lightest charged particles that traverses the ATLAS detector. As described in chapter , the electron mainly deposits energy by radiating off photons through bremsstrahlung and transition radiation. However, it does also ionize. Therefore, as visualized in figure 16, the electron bends in the magnetic field, leaving hits in the ID and the TRT, and terminates in the ECAL in an electromagnetic shower, depositing almost all ($\approx$ 99%) of its energy. This is a clear signature and deviations are mostly due to the electron bremming a photon before exiting the ID, which will slightly change the shower shape.

Muons travel through the ATLAS detector mostly as a MIP, as described in chapter . This means that only highly energetic muons will deposit a significant amount of energy through other processes than ionization. It, therefore, travels all the way through the detector leaving a varying amount energy deposits in the ID, ECAL, Hadronic Calorimeter (HCAL), and MS. Measurement of the momentum of the muon is therefore left to the bending of solenoid magnet in the ID and the toroidal magnets in the MS.

Neutrinos will absolutely not interact with the ATLAS detector and was it not for the conservation laws of physics, they would remain unmeasured. As previously described, a vertex of interaction in particle physics will be accompanied by multiple laws of conversion. By accounting for the whole detector, a measure of the missing transverse energy can be made. This allows the detector and accompanying reconstruction software to determine the energy and direction of neutrinos. Although it certainly does not allow the same efficiency as for electrons, due to compounding effects on resolutions.

Hadronization Jets behave, as described in , in a quite chaotic manner. Due to their long mean free path they deposit most of their energy in the HCAL but can also start an electromagnetic shower in the ECAL. This, along with the large jet production cross section (figure 7) and the fact that jets contain an endless list of particles makes them extremely versatile background objects. Jet fakes are a large background for almost all identification in ATLAS.

Photons in the ATLAS Detector

Light in the ATLAS detector has a special role in this thesis and will therefore be described in further detail. There are several important features that make photons rather complicated. The photon is a massless, uncharged particle and, as described in chapter , it does not interact with the tracking part of the detector, nor the magnetic field. It terminates by decaying into an e^+e^- pair through pair production and producing an electromagnetic shower in the ECAL. However the exact same mechanism with which it showers is what complicates

them. A photon will often (about 50% of the time) pair produce before reaching the ECAL. This leads to two very collimated electrons, that affect the shower shape in various ways, depending on when the conversion occurs. This further muddies the water between an electron and a photon in the detector, since the deposit in the ECAL will be originating from actual electrons. This is referred to as a conversion and the resulting object is a converted photon. Photons that does not convert before the ECAL are referred to as unconverted photons. There are several different types of conversions, differentiated by the location and quality of the hits left in the ID. A list of the conversion types along with a short description is found in table 2. The probability of converting increases with material and is not uniform in η. There is a general trend of higher conversion rates as η increases ("Before Accordion" on figure 12). As mentioned this complicates photon identification significantly, the foundation of which is the shower shapes, which are substantially impacted by conversions. Additionally there is now further information to be gained from the tracks.

Conversion Type	Description
0	Unconverted photons
1	Single track associated with the cluster, with hits in the Silicon pixel detector
2	Single track associated with the cluster, with only hits in the TRT
3	Two tracks associated with the cluster, with hits in the Si
4	Two tracks associated with the cluster, with only hits in the TRT
5	Two tracks associated with the cluster, with only one having hits in the Si.

Table 2: The photon conversion types and their description.

Electron and Photon Reconstruction in ATLAS

The reconstruction of electrons and photons in ATLAS is done with an algorithm that takes the superclusters built in the calorimeter and the tracks from the ID as input. The task at hand, simplified, is then to match the clusters with a track, resulting in an electron, match the clusters with a conversion vertex, resulting in a converted photon or find no matched tracks for the clusters, resulting in photons. The process to select these clusters and tracks will be described below.

Topo-Clusters[**Lampl:1099735**] are reconstructed from proto-clusters. They are initiated by a requirement for cells in layer 2 or 3 of the ECAL to pass a noise threshold of at least 4 times the expected noise[11]. Neighboring cells that pass a noise threshold of at least 2 times the expected noise are added to the proto-cluster. If two clusters share this cell, they are merged. Finally neighboring cells passing a noise threshold of at least 0 times the expected noise are added. From here, proto-clusters can be split given a criteria, that indicates it is in fact two separate deposits. Electron and photon reconstruction only uses topo-clusters that have a significant fraction

[11] This is calculated from electronic noise and pileup noise

of their deposit in the ECAL, given this is their expected deposit.

Tracks[**collaboration2019electron**] are reconstructed from the hits in the ID and rely on pattern recognition[**Cornelissen:1020106**]. Initially, track seeds are created from three space-points in the pixel or microstrip subdetectors of the ID. From here, three steps are taken: pattern recognition, ambiguity resolution and TRT extension. Hereafter a series of Kalman filters refits the tracks to better account for energy lost while traversing the ID. This is specifically done in Regions of Interest (ROIs) chosen by EM clusters that have a significant fraction of their deposit in the ECAL. The tracks are then matched to clusters with requirements on the difference in η and ϕ measurements of the second ECAL layer and the track extended into this layer. If this leads to multiple tracks being associated with the same cluster, they are ranked according to smallest ΔR.

Conversion Vertices[**Aaboud_2019**] are reconstructed by attempting to match two opposite charge tracks into a vertex consistent with the one of a massless particle, leading to two-track conversions. Single-track conversion are made from tracks that does not have hits in the inner most layers of the ID. This is done with tracks that are loosely matched to clusters. Silicon tracks[12] used to create a conversion vertex are required to have a high probability of being an electron track, as determined by the TRT. This requirement is higher for TRT tracks and even higher for tracks used to create single conversions. They are then matched to topo-clusters. Given multiple candidates, the preferred track will be the one of the highest quality, denoted by location of hits and radius of conversion.

[12] Hits in the pixel or microstrip subdetectors

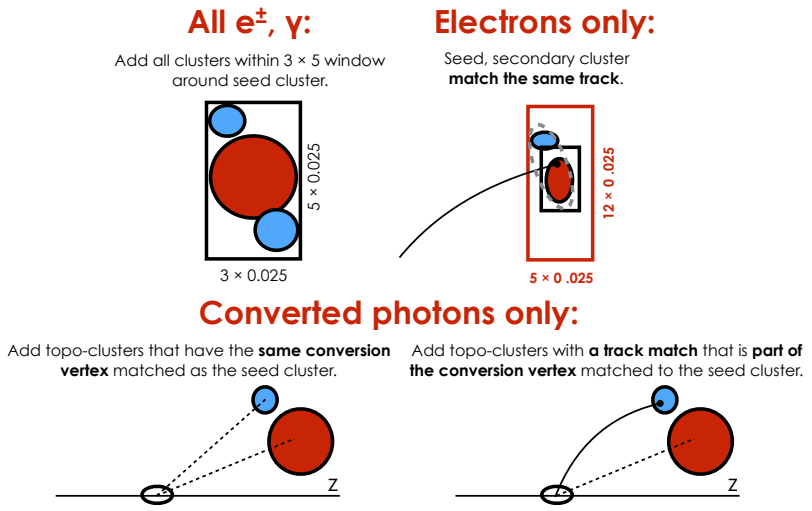


Figure 17: Diagram of the supercluster algorithm for electrons and photons. Figure taken from [**ATL-PHYS-PUB-2017-022**]

As mentioned, the Supercluster reconstruction takes the three described entities as input and then proceeds to process electrons and photons separately. A diagram of the algorithm is shown in figure 17. The top left part shows the initial step, with seed clusters shown

in red and satellite clusters in blue. A cluster is then deemed to be an electron supercluster seed if it has an associated track with at least 4 hits in the silicon detectors and a deposit of at least $1GeV$. Photon supercluster seeds are just required a deposit greater than $1.5GeV$. Satellite clusters are then added by requiring them to be near the seed cluster and further track requirements are passed for satellite electron supercluster seeds and photon supercluster seeds with an associated conversion vertex. The result is electron and photon superclusters, whose position and energy are recalibrated and tracks (conversion vertices) are matched to electron (photon) superclusters in the same way as the topo-cluster matching described above. A supercluster can end up being selected as both a photon and electron supercluster seed. It is not viable (For reasons of memory) to keep the object is both containers and ambiguity resolution is done. This basically selects clusters with a good matched track and no conversion vertex as electrons and clusters with no good track matched as photons. However, where no resolution can be made, the object is saved as both an electron and a photon and marked as ambiguous. From these final photons and electrons variables can now be calculated and used for identification. This is needed because many objects will be reconstructed as a given object, when in truth they are something entirely different. In the following section the electron and photon identification deployed by ATLAS will be described. Given that the aim of this thesis is to provide an alternative and evaluation will, therefore, involve direct comparison with the ATLAS identification it is important to understand how it works.

Electron and Photon Identification in ATLAS

The complexity of the problem can be described as follows. As visualized by figure 16, one can easily imagine being able to point to the hits in the ID and the deposit in the ECAL and then determine whether the object is a photon, electron or something different. Now add on the objects from 50 additional interactions, and the problem increase in difficulty. Now imagine having to do this correctly for more than 90% of electrons while only making mistakes once every 100000 objects. Given infinite time, surely, this is possible. Realizing that this task has to be done continually for billions of events a year, the problem starts to show its true complexity. While every given task is not hard, the sheer number of configurations, frequency, and required precision makes it extremely challenging.

Luckily ATLAS does not employ people to discriminate particles, they employ computers, many, many computers. In order to ask a computer to solve this problem, the first step is to translate the information about the objects, be it a photon or electron, to the computer. A great amount of work has been put into developing, simulating and properly reconstructing a list of variables used by ATLAS for electron and photon identification, this list along with a short description and their use cases is shown in table 3. A schematic rep-

Name	Description	Usage
Hadronic Leakage		
R_{had_1}	Ratio of E_T in the first sampling of the hadronic calorimeter of E_T of the EM cluster (for $\|\eta\| < 0.8$ and $\|\eta\| > 1.37$)	e/γ
R_{had}	Same as R_{had_1} (for $0.8 < \|\eta\| < 1.37$)	e/γ
EM Third Layer		
f_3	Ratio of the energy in the third layer to the total energy in the EM	e
EM Middle Layer		
R_η	Ratio between the sum of the energies of the cells contained in a 3×7 ($\eta \times \phi$) rectangle and the sum of energies in a 7×7 rectangle	e/γ
w_{η^2}	Lateral shower width, $\sqrt{(\sum E_i \eta_i^2)/(\sum E_i) - ((\sum E_i \eta_i)/(\sum E_i))^2}$, calculated within a window of 3×5 cells	e/γ
R_ϕ	Ratio between the sum of the energies of the cells contained in a 3×3 ($\eta \times \phi$) rectangle and the sum of energies in a 3×7 rectangle	e/γ
EM Strip Layer		
w_{s3}	Lateral shower width, $\sqrt{(\sum E_i (i - i_{max})^2)/(\sum E_i)}$, where i runs over all strips in a window of 3 strips aounrd the highest-energy strip, of index i_{max}	γ
w_{stot}	Same as w_{s3}, however the the windows is of size $\Delta\eta \approx 0.0625$ and i_{max}	e/γ
f_{side}	Fraction of energy outside core of 3 central strips, but within 7 strips	γ
ΔE_s	Difference between the energy of the strip associated with the second maximum in the strip layer, and the energy reconstructed in the strip with the minimal value found between the first and second maxima	γ
E_{ratio}	Ratio of the energy difference between the maximum energy deposit and the energy deposit in a secondary maximum in the cluster to the sum of these energies	e/γ
f_1	Ratio of the energy measured in the first samlping of the EM calorimeter to the total energy of the EM cluster	e/γ
Track Conditions		
n_{Blayer}	Number of hits in innermost pixel layer	e
n_{Pixel}	Number of hits in the pixel detector	e
n_{Si}	Total number of hits in the pixel and SCT detectors	e
d_0	Transverse impact parameter relative to the beam-line	e
$\|d_0/\sigma(d_0)\|$	Significance of d_0	e
$\delta p/p$	Momentum lost by the track between the perigee and the last measurement point divided by the momentum at perigee	e
$ePorbabilityHT$	Likelihood probability bsaed on transition radiation in the TRT	e
$\Delta\eta_1$	$\Delta\eta$ between the cluster position in the first layer matching and the extrapolated track	e
$\Delta\phi_{res}$	$\Delta\phi$ between the cluster position in the second layer of the EM calorimeter and the momentum-rescaled track, extrapolated from the perigee, times the charge q	e
E/p	Ratio of the cluster energy to the track momentum	e

Table 3: Discrimating variables used by ATLAS for electron and photon identifcation. The usage listed in the last column refers to the type of identificaiton the variable is applied. Highly inspired by internal ATLAS documentation.

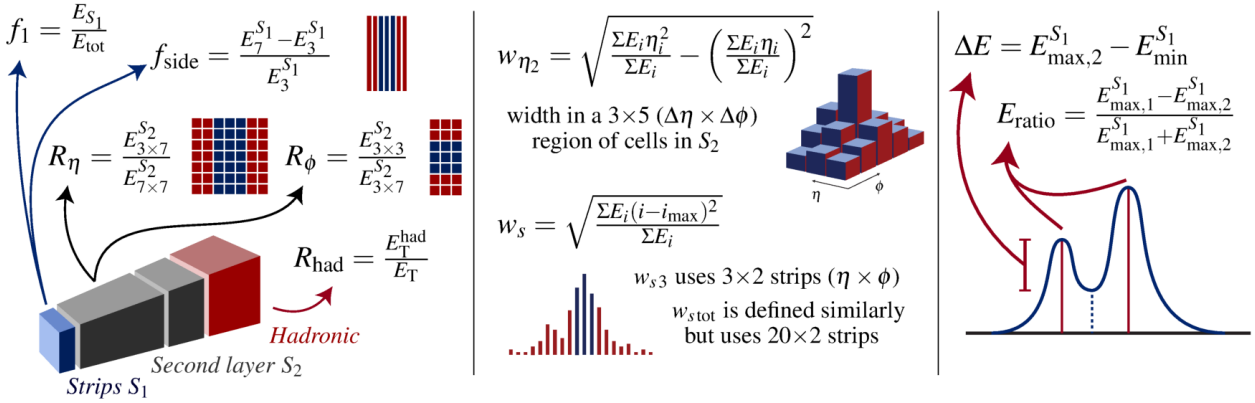


Figure 18: Schematic representation of the discrimanating variables used for the Photon identification, some of which are also used for electron idenfication. Figure taken from **[Aaboud_2019]**

resentation of the discriminating variables used for photon identification (Some shared by electron ID) is shown in figure 18, to give some context to the descriptions from the table. The foundation of the identification is the shape of the showers in the ECAL. The knowledge applied is the fact that electrons, photons and jets all deposit their energy in the ECAL differently. Therefore, designing variables that describe the shape of these deposits will allow a computer to discriminate between the object types by looking at distributions of these shower shape variables and applying a selection based on them. The way this selection is applied is very different between electrons and photons. They will therefore be described separately.

Electron Identification The electron identification aims to select isolated and prompt electrons in a range of $|\eta| < 2.47$ [13] and uses the variables, listed in table 3 with usage of "*e*", in different ways. All but the following variables: w_{stot}, n_{Blayer}, n_{Pixel}, n_{Si}, and E/p are used to develop a likelihood (LH) discriminant. It is based on the product for signal, L_S, and for background, L_B, of a probability density function (PDF)[14] for each variable, P:

$$L_{S(B)}(\boldsymbol{x}) = \prod_{i=1}^{n} P_{S(B),i}(x_i) \tag{4}$$

Where L_S and L_B refers to the probability of being signal and background, respectively. n represents the number of variables (and PDFs made). $P_{S,i}(x_i)$ and $P_{B,i}(x_i)$ is the value the PDFs take for the variable i at value x_i for signal and background respectively. The likelihood discriminant, d_L, is defined as:

$$d_L = \frac{1}{\tau} ln(\frac{L_S}{L_B}) \tag{5}$$

where $\tau = 15$. The signal PDFs are derived from $Z \to ee$ and $J/\psi \to ee$ for E_T above and below $15GeV$, respectively, using an involved tag and probe method that is described in **[collaboration2019electron]**. The LH pdfs are determined in bins of $|\eta|$ and E_T (9 $|\eta|$ bins and 7 E_T bins) to combat the significant differences in shower shapes across

[13] excluding the crack

[14] PDF will in this section refer to this instead of Parton Distribution Function

referred to as

these variables. Four working points[15] VeryLoose, Loose, Medium, and Tight are defined based on increasing thresholds in d_L. This discriminant is chosen in the same bins of $|\eta|$ and in a finer binning in E_T (12 bins). The variables unused by the LH discriminant comes into play for the Loose, Medium, and Tight working points, by requiring $n_{Pixel} \geq 2$ and $n_{Pixel} + n_{Si} \geq 7$, and $n_{Blayer} \geq 1$ for Medium and Tight. Many more nuances are dealt with, for instance the discriminant slightly altered as a function of pileup, and w_{stot} and E/p are used to recover lost efficiency for the tight working point at high energy.

Photon Identification The main goal of the photon identification is to select prompt, isolated photons in a range of $|\eta| < 2.37$ [16], while reducing background from hadronic jets that has a high fake rate due to actual photons from for instance π^0s. The selections deployed by ATLAS to select photons are cut based selections, which rely on rectangular cuts on variable distributions. The variables used were introduced in table 3 and described further in the schematic of figure 18. The two main working points of the photon identification are Loose and Tight. The Loose working points is made up of rect-

[16] Excluding the crack

$\|\eta\|$ range	$0.0-0.6$	$0.6-0.8$	$0.8-1.15$	$1.15-1.37$	$1.52-1.81$	$1.81-2.01$	$2.01-2.37$
R_{had}	0.0606	0.0524	0.0648	0.0491	0.0479	0.0651	0.0620
R_{eta}	0.9175	0.8999	0.9109	0.8912	0.8922	0.9221	0.8903
w_{η^2}	0.0129	0.0142	0.0136	0.0139	0.0152	0.0128	0.0125
$E277$				0.1			

Table 4: The Loose ID cuts for both unconverted and converted photons. Table taken from **[brendlinger_s_2019]**

angular cuts on a few variables, the variables and the corresponding cuts are found in table 4. The cuts vary in bins of $|\eta|$ to account for changing shower shapes. The Loose working point is a rather simplistic selection, designed to select $\sim$ 99% of signal while lowering background with around a factor 1000**[brendlinger_s_2019]**. It is mostly used for triggering and pre-selection and is in fact used for pre-selection in the optimization of the Tight working point. This allows for the use of loose triggers and makes the Tight working point a subset of the Loose working point, which is ideal. The Tight working point is optimized in the same bins of $|\eta|$ as the Loose, however, it is also optimized in bins of E_T and separately for unconverted and converted photons. The bins in $E_T[GeV]$ are the following: $25-30-35-40-50-60-80-100-125-150-175-250-500-1500$. The ten variables described in figure 18 are used for the Tight working point. Fudge factors are applied to the shower shapes in order to combat MC mismodelling. The rectangular cuts on the 10 variables are optimized using the ROOT framework for multivariate analysis, TMVA**[hoecker2007tmva]**. This selects running cuts, which vary in efficiency to select signal, ϵ_{sig}, and background, ϵ_{bkg}. Calculated as the fraction of signal and background objects that pass the cuts out of the total number of signal and background objects. The optimal

cuts are then selected by optimizing the significance:

$$S = \epsilon_{sig} / \sqrt{\epsilon_{bkg}} \tag{6}$$

The optimization of cuts is done in bins of $|\eta|$ and E_T. In order to obtain enough statistics for this, an effort has to be made to use a data-driven background selection at high and low energies, optimizing the cut on MC signal against data background. A summary of the optimization in different bins of E_T can be found in table 5.

Signal	Trained Against	E_T Range
MC $Z\gamma$	Data (without cuts on M_{ll} and $M_{ll\gamma}$)	$10 < E_T < 25GeV$
MC $\gamma - jet$	MC dijets	$25 < E_T < 100GeV$
MC $\gamma - jet$	Data	$100 < E_T < 1500GeV$

Table 5: The different choices for signal and background for optimization of the Tight working point. The choice to use a data-driven background selection is to combat lack of statistics in edge bins.

Isolation

This thesis will also train isolation models for reasons related to data training and the term isolation will be used frequently. In ATLAS isolation refers to the particle being relative undisturbed by other particles. The reason this is important is because the particles from hard scatter events, such as $Z \to ll\gamma$ or $H \to \gamma\gamma$, are expected to be relatively isolated, since the hard interaction only produces those respective photons and electrons. The goal here is to distinguish particles from hard interactions from the ones produced in soft interactions, which will contain both photons and electrons. As mentioned, the most commonly produced particles are pions, which produce plenty of photons and electrons. These are actual electrons and photons, however they will most likely not be isolated.

Isolation is measured both in the tracker and in the calorimeter. The isolation of an object is calculated in the tracker by removing the track belonging to the object and summing up the momentum of the remaining tracks in cones of $\Delta R < 0.2$, 0.3, 0.4. Leading to variables *ptcone*20, *ptcone*30, and *ptcone*40. Doing similarly for the calorimeter deposits (removing the cluster) and acquires the variables: *etcone*20, *etcone*30, and *etcone*40. The isolation models trained will use a mixture of these variables along with $\langle\mu\rangle$, $|\eta|$, and E_T.

The Data and Datasets

This chapter will describe the event selections used to produce MC and DATA files for training and evaluation of particle identification models. The resulting photon datasets will also be visualized by plotting distributions of $\langle\mu\rangle$, η, and E_T.

The work in this thesis touches different datasets and each is the result of a unique selection, each of which will be explained in this section. The selections are the following: the truth matched MC selection for photons, a full Tag and Probe ($T\&P$) framework with signal and background selections for electrons in data, and a $TT\&P$ selection for photons in data. All of which will be described separately. However, first some general points about the framework will be touched upon. This section hopes to provide an insight into the approach by describing each step and the decisions made along the way.

There is a long list of xAOD productions available from EGamma (table 6). These were heavily used for inspiration, however given the nature of the overall project, they would not be sufficient. This is mainly due to the wish of including the images of the calorimeter deposits. An in-house DxAOD production was written by Lukas, which is a mix of $EGAM1$, $EGAM2$, $EGAM3$, $EGAM4$, and $EGAM7$, with cell decoration added on top. This is done in order to encompass ntuple production for both electrons and photons. This was used for photon MC dataset and the electon DATA dataset. However, due to time limits, the EGamma derivations were used for the photon DATA dataset.

Deriviation	Description
EGAM1	$Z \to ee$ (central electrons)
EGAM2	$J/psi \to ee$
EGAM3	$Z \to ee\gamma$, $Z \to eee$
EGAM4	$Z \to \mu\mu\gamma$, $Z \to \mu\mu e$
EGAM5	$W \to e\nu$
EGAM6	$Z \to$ (looser than EGAM1)
EGAM7	Inclusive electrons
EGAM8	$Z \to ee$ (At least one forward e)
EGAM9	Bootstrap for photon trigger eff.

Table 6: The 9 different EGamma deriviation frameworks with a short description

The Electron Dataset (DATA) For the electron dataset, the selection is outlined by the work flow in figure 19. This selection was written by another student and aims to provide a versatile selection that picks up as many signal candidates as is realistic, while also selecting as much relevant background as possible. Given the peripheral use of electrons for the work in this thesis, the selection will not be covered in great detail. However, for a much more detailed description see **[ehrke_machine_2019]**. The triggers required in the selections outlined in the work flow can be found in appendix . As indicated by the flow diagram, one first selects an assortment of tags and probes, namely tag and probe electrons for $Z \to ee$ and $J/\psi \to ee$ $T\&P$, tag muons and background electrons for background selection. The output of this pre-selection is then sent through an event selection that

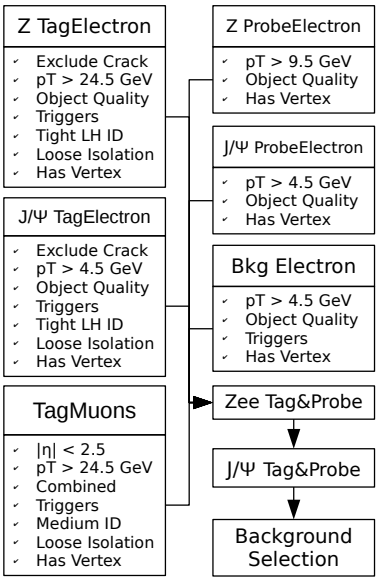


Figure 19: Work flow of the selection used for electron dataset. The selection is made for the work of another student**[ehrke_machine_2019]**

will briefly be outlined.

Before any event selection is done, event cleaning and a requirement for there to only be either a tag electron or a tag muon in the event is passed. The $Z \rightarrow ee$ *T&P* is then done, which involves matching Z tag electrons with probe electrons, required to have $\Delta R > 0.4$ and $m_{ee} < 50GeV$. If the event contains one Opposite Sign (OS) pair it is kept as signal, if it contains more than one the event is vetoed[17]. If none, all possible Same Sign (SS) pairs are saved as background and the selection moves on. Next is a μe selection, which attempts to match tag muons and Z tag electrons with the same requirements as the $Z \rightarrow ee$ *T&P*. This is background selection and every pair found is saved as such. Next in line is the $J/\psi \rightarrow ee$ *T&P*, which requires $\Delta R < 0.1$ and $1.5GeV < m_{ee} < 5.0GeV$. A requirement on the pseudo proper lifetime of the J/ψ is also required. Only if there is one OS pair it is saved as signal, otherwise all background particles are saved, while vetoing the Z- and W-mass regions to reduce signal contamination of background.

[17] Nothing is saved

Table 7: AODs used for the Electron Dataset, the in-house DxAOD production was run on these after which the described electron event selection was run.

AODs for the Electron Dataset
data17_13TeV.00327862.physics_Main.merge.AOD.r10203_p3399
data17_13TeV.00327582.physics_Main.merge.AOD.r10203_p3399
data17_13TeV.00330160.physics_Main.merge.AOD.r10203_p3399
data17_13TeV.00327761.physics_Main.merge.AOD.r10203_p3399
data17_13TeV.00327490.physics_Main.merge.AOD.r10203_p3399
data17_13TeV.00326446.physics_Main.merge.AOD.r10250_p3399
data17_13TeV.00328374.physics_Main.merge.AOD.r10203_p3399
data17_13TeV.00327265.physics_Main.merge.AOD.r10250_p3399
data17_13TeV.00327636.physics_Main.merge.AOD.r10203_p3399
data17_13TeV.00326657.physics_Main.merge.AOD.r10250_p3399
data17_13TeV.00328099.physics_Main.merge.AOD.r10203_p3399
data17_13TeV.00330101.physics_Main.merge.AOD.r10203_p3399
data17_13TeV.00326551.physics_Main.merge.AOD.r10250_p3399
data17_13TeV.00330079.physics_Main.merge.AOD.r10203_p3399
data17_13TeV.00326945.physics_Main.merge.AOD.r10250_p3399
data17_13TeV.00327057.physics_Main.merge.AOD.r10250_p3399
data17_13TeV.00327745.physics_Main.merge.AOD.r10203_p3399
data17_13TeV.00325713.physics_Main.merge.AOD.r10260_p3399
data17_13TeV.00329829.physics_Main.merge.AOD.r10203_p3399

It is worth nothing that the J/ψ selected signal is not used for the work in this thesis. The dataset used contains the signal selected by the $Z \rightarrow ee$ *T&P* and all the selected background particles. The AODs used to create the dataset are found in table 7. This is a subset of data17 and was deemed a sizable dataset for the need of this work. However, more statistics are available.

The Photon Dataset (MC) In MC, the photon selection is unfortunately rather simple. It is a loop over everything reconstructed as a photon (the photon container), with truth matched photons being marked as signal and everything else as background, only applying

a modest $p_T > 4.5GeV$. This selection is then run on a long list of MC DAODs, with the full list found in appendix . Taking advantage of the photon container being full of particles that to the reconstruction behaves much like a photon provides a readily available dataset with incredible statistics. However, it does not necessarily provide a background distribution similar to one relevant for selections one might make in data. Performance on resonances, as well as in data, might be limited. This also limits comparability with the ATLAS cuts, since they are optimized on a different background composition as described in . By allowing a model to see everything that is present in the photon container it has potential to generalize across many selections both in MC and data, due to the fact that most selections aiming to select photons will have its foundation in the photon container. The hope is simply that the model will be able to encompass this massive dataset and in turn minimize the effect of the aforementioned worries.

Label	Total [%]	Composition [%] and Description	
Signal	7.91	100.0	Isolated Photons
Background	92.09	70.31	Background Photons
		19.30	Non truthmatched Objects
		8.23	Hadrons
		1.36	Electrons
		0.73	Non Isolated Photons

Table 8: The composition of the photon MC dataset, separated in signal and background

The composition of the labels in the MC dataset is found in table 8 with a full table of the statistics separated in bins of $\langle\mu\rangle$, $|\eta|$, and E_T seen in table 9. This clearly shows the vast amount of statistics available in the MC sample.

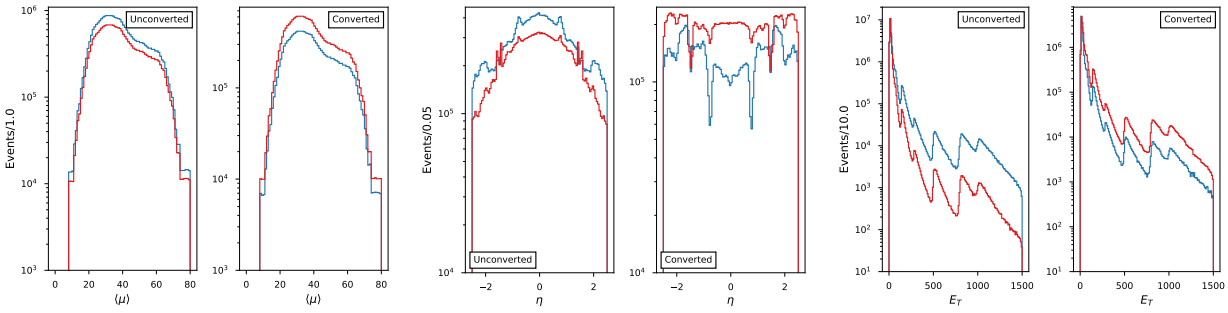


Figure 20: The distributions of $\langle\mu\rangle$, η, and E_T for the converted and unconverted channels. The blue histogram is signal, while the red is background. Weights are applied, they will be described in ML methods

The Photon Dataset (DATA) In DATA, the photon selection is an expansion of a classic $T\&P$ framework for selecting electrons from $Z \to ee$, made by adding a 2nd tag. Similarly to the $T\&P$, we intend to select probes as unbiased as possible, which in this case includes not applying any cuts to the invariant mass of the lepton pair, neither with nor without the photon. This is possible by applying identification, isolation and E_T requirements on both tags. The philosophy remains the same, one makes sure the tags are truly electrons and then looks around for something that resembles a photon, which for

Dataset	Train				Validation				Testing			
Bin	Unconverted		Converted		Unconverted		Converted		Unconverted		Converted	
$\langle\mu\rangle$	N_{sig}	N_{bkg}	N_{sig}	N_{bkg}	N_{sig}	N_{bkg}	N_{sig}	N_{bkg}	N_{sig}	N_{bkg}	N_{sig}	N_{bkg}
$10-20$	221523	2085936	104534	866416	26988	261819	12825	107776	28525	260794	13375	108242
$20-30$	1088626	11250204	515398	4468486	132996	1406987	62707	556350	138570	1398015	65294	555142
$30-35$	689688	7725611	329568	3024389	84135	966782	40103	377432	88301	961348	42062	377960
$35-40$	631071	7501886	304245	2937456	76721	937617	36952	367597	80522	932559	38435	365905
$40-50$	836832	10864783	409662	4308417	102346	1360581	49892	538008	106691	1351400	52358	535068
$50-60$	606180	9021544	302626	3645773	73834	1130961	36945	457878	77523	1123568	38427	453641
$60-80$	393805	6760385	197863	2784484	48305	845801	23965	348626	50457	839879	25368	346348
Total	4467725	55210349	2163896	22035421	545325	6910548	263389	2753667	570589	6867563	275319	2742306
	83877391				10472929				10455777			
$\lvert\eta\rvert$	Train				Validation				Testing			
$0.0-0.6$	1513765	21818892	426416	9897235	184625	2731508	51641	1238086	194110	2715372	54226	1233113
$0.6-0.8$	487688	6408049	109754	2083833	59584	802663	13271	258980	62671	797782	14099	259115
$0.8-1.2$	780123	10419474	349955	3075695	94066	1304551	42176	385079	99969	1296444	44829	382815
$1.2-1.37$	287480	3490612	164292	1032146	34948	437205	19888	128459	36641	433904	20947	128858
$1.37-1.52$	219064	2903034	129201	674850	27187	362835	15860	84494	27690	361064	16553	83605
$1.52-1.81$	361830	4943161	346039	1656789	44304	618343	42214	207704	45833	614258	44189	205638
$1.81-2.37$	685811	4525576	523084	2964367	84267	565036	64010	369226	87158	561951	65941	368376
Total	4335761	54508798	2048741	21384915	528981	6822141	249060	2672028	554072	6780775	260784	2661520
	82278215				10272210				10257151			
$E_T[GeV]$	Training				Validation				Testing			
$4.5-10.0$	441382	40979008	132596	7248790	54922	5130929	16462	907322	54940	5095563	16504	900269
$10.0-20.0$	1679803	9340368	779954	2909591	209861	1167668	97273	363301	209561	1163291	97328	361132
$20.0-30.0$	764753	1858121	436819	1339734	95108	231834	53905	167397	95869	231766	54417	166525
$30.0-40.0$	365271	736091	217474	847470	44226	92321	26735	106090	46867	91458	27698	105994
$40.0-50.0$	180977	382276	103585	573158	20075	47755	11799	71553	24832	47838	13842	71190
$50.0-60.0$	133419	219174	68184	391654	13488	27491	6959	48996	20061	27050	9969	48931
$60.0-80.0$	201762	251358	102175	556497	21420	31469	10934	69119	29298	31483	14524	69081
$80.0-100.0$	108372	144089	56139	411097	12513	18015	6627	51705	14483	17934	7584	51429
$100.0-150.0$	127737	198925	65488	761296	15611	25124	7704	94882	16349	24628	8449	94827
$150.0-250.0$	176401	182835	89347	1059136	21900	22916	11079	132168	22287	22807	11070	132528
$250.0-1500.0$	192613	334268	87144	4992665	24057	41733	10750	623289	24164	41397	10788	623716
Total	4372490	54626513	2138905	21091088	533181	6837255	260227	2635822	558711	6795215	272173	2625622
	82228996				10266485				10251721			

Table 9: The statistics of the MC photon dataset, separated into bins of $\langle\mu\rangle$, $|\eta|$, and E_T. The first of the two totals are the sum in the columns

this selection is quite loosely defined.

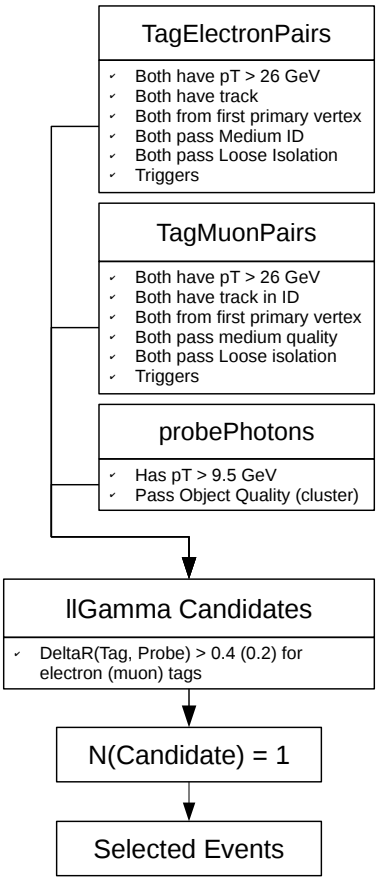


Figure 21: Flow diagram describing the $TT\&P$ selection, described in detail in text.

A flow diagram of the $TT\&P$ method is seen in figure 21. Initially, it selects lepton pair candidates by running two separate selections, one on the electron container and one on the muon container. Each selection attempts to match every possible pair with the selection criteria listed in figure 21. A small selection is run on the photon container to select probes. They are required a minimal p_T of $9.5GeV$ and an acceptable object quality, by requiring it to not be a bad cluster photon. This is an EGamma definition and it checks for faulty, missing or otherwise poor readouts from cells in the cluster.

These selected lepton pairs and probe photons are then matched with the only requirements being a cone distance ΔR (remember $\Delta R \equiv \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$) between each individual tag and the probe to be higher than 0.2 (0.4) for electron (muon) tags. Additionally there can only be one candidate in the event. It is possible to change this to selecting the candidate with the highest probe p_T. However, due to their rarely being multiple candidates in the event, this was not added before large scale production. If one is to apply looser requirements on the tag pairs, one should probably consider such an addition. Due to time constrains the dataset as created using the

Table 10: The DAODs supplied by E/γ used for creation of the photon dataset

DAODs for $Z \to ee\gamma$
data17_13TeV.periodI.physics_Main.PhysCont.- DAOD_EGAM3.repro21_v01
data17_13TeV.periodC.physics_Main.PhysCont.- DAOD_EGAM3.repro21_v01
data17_13TeV.periodF.physics_Main.PhysCont.- DAOD_EGAM3.repro21_v01
data17_13TeV.periodD.physics_Main.PhysCont.- DAOD_EGAM3.repro21_v01
data17_13TeV.periodK.physics_Main.PhysCont.- DAOD_EGAM3.repro21_v02
data17_13TeV.periodE.physics_Main.PhysCont.- DAOD_EGAM3.repro21_v01
data17_13TeV.periodH.physics_Main.PhysCont.- DAOD_EGAM3.repro21_v01
data17_13TeV.periodB.physics_Main.PhysCont.- DAOD_EGAM3.repro21_v01
DAODs for $Z \to \mu\mu\gamma$
data17_13TeV.periodAllYear.physics_Main.PhysCont.DAOD_EGAM4.grp17_v01_p

Egamma DAODs listed in table 10, they should encompass all of data17.

The Higgs Selection was built to simply run on $H \to \gamma\gamma$ MC DAODs. Due to the interests of another student, the HIGG1D1 derivation framework was expanded to include images and run on 300mc16_13TeV-.343981.PowhegPythia8EvtGen_NNLOPS_nnlo_30_ggH125_gamgam.-merge.AOD.e5607_e5984_s3126_r10201_r10210 AOD files. The selec-

tion was then run on the resulting DAODs. It is event-based and is built to perform a simple, loose pre-selection, in order to get all reasonable di-photon pairs through. The MC file contains pileup with some of it making it through the selection. Nevertheless, a similar event selection as the one applied by ATLAS, described in chapter , is applied later and this leaves only actual diphoton pairs from Higgs decays.

The pre-selection applied is described in table 11.

Higgs MC selection
Overlap Removal
$E_T < 4.5GeV$
Loose Isolation
Loose Identification

Table 11: The requirements of the eventbased ntuple production built to run on $H \rightarrow \gamma\gamma$ DAODs

Variables

In the mindset of particle identification, the vast majority of information available, about a given object in the ATLAS framework, provides little discrimination power. In general there is a preference for a limited variable list in EGamma, which is echoed in this work. This is mainly due to the upkeep needed, should a long list of extra variables be included into an extended framework for particle identification. The mentality quickly became for the foundation to be the understood and well simulated variables already in use by the current framework. New variables were put under scrutiny. In the end an extended list of variables has been used in this work and they are shown, along with a short description in table 12. They are introduced now for reference. It is worth noting that $\langle\mu\rangle$ as introduced here and used throughout the thesis refers to the number of interactions calculated over LumiBlocks. As to not confuse it with the number quoted in figure 6 in chapter . The number used refers to the distributions shown in the very same figure. Additionally, the number of reconstructed vertices used by the electron training is a measure of much the same thing, inferred from the amount of vertices available after event reconstruction, it should make for a better measure for a given event.

Variable name	Description
Photon variables	
$maxEcell\_time$	The time value of the cell in the cluster with the highest energy deposit
$maxEcell\_energy$	The energy value of the cell in the cluster with the highest energy deposit
$r33over37allcalo$	The ratio of the energy deposits in the 3×3 over 3×7 in the whole ECAL, centered around the cluster
$\langle \mu \rangle$	The average number of interactions in the LumiBlock the event was recorded
Electron variables	
n_{tracks}	The number of tracks associated with the cluster
$E7 \times 11\_Lr3$	Energy deposited in 7×11 cells centered around the cluster in layer 3
$E3 \times 5\_Lr1$	Energy deposited in 3×5 cells centered around the cluster in layer 1
$E3 \times 5\_Lr0$	Energy deposited in 3×5 cells centered around the cluster in the pre-sampler
$E7 \times 11\_Lr2$	Energy deposited in 7×11 cells centered around the cluster in layer 2
$fracs1$	f_side from figure 18
p_T track	The momentum of the track, measured by the ID
$\Delta\eta_2$	The difference in the measurement of η from the track and calorimeter layer 2
ambiguityType	Whether the object is marked as ambiguous or not, introduced in chapter
f_1 core	The energy deposited ± 3 strips in η of layer 1, centered around the highest cell, divided by the total energy of the calorimeter
$\Delta\eta_0$	The difference in the measurement of η from the track and calorimeter layer 0
n_{vertex} reco	Number of reconstructed vertices for the event
Electron variables	
$core57cellsEnergyCorrection$	An energy correction, calculated from the 5×7 cells in the calorimeter, centered around the cluster
E_T calo	The energy as reconstructed by the calorimeter
η	The pseudo rapidity of the particle, as introduced in chapter
Isolation variables	

Table 12: The variables used that were not introduced in the ATLAS chapter. The shared variables are used indirectly by ATLAS and has been described here for clarity

The Machine Learning Methods

The majority of the work done in this thesis mainly involves machine learning. The physics knowledge gained was mostly applied in order to better design the models as well as interpret their outputs. This has proven to be a fruitful methodology when approaching physics problems with an intend to apply machine learning methods as seen by trackML challenge[**noauthor_trackml_nodate**]. Even the best machine learning has a tendency to falter when proper and thorough physical understanding of the processes and parameters involved is not applied. Nevertheless, it is equally important to understand the wide range of possible machine learning methods and consider their possible strengths and weaknesses when tackling a certain problem. Therefore, this chapter will serve as an introduction to and guidebook of the methods used in this thesis. It will clearly state the approach, the challenges and the neat tools applied to further increase performance, ease optimization and more.

General Method

In the general, the method applied attempt to take advantage of the incredible statistic of HEP, in the best possible way. 20% of data is set aside for validation and evaluation, half for each. The rest is used exclusively for training. The important of validation cannot be understated. The large amount of statistics provide for rather long training times and applying early stopping make sure this training time is spent in the best possible way. For the tree based methods a preprocessing step was added. The $\langle\mu\rangle$, $|\eta|$, and E_T distributions were reweighted to be equal between signal and background. This is done using a tool called Gradient Boosted Reweighter (GBR)[**noauthor_arogozhnikov_nodate**]. This allows for reweighting in three dimensions with a much better result than bin by bin reweighting[**ehrke_machine_2019**]. The motivation behind the reweighting is to minimize the correlations between the outputs and these variables. The goal is for them to have discriminating power and since the three based methods apply cuts on distributions, this should be true after reweighting. The degree to which the reweighting works is visualized by figure 22. The resulting weights will be applied while training and evaluation.

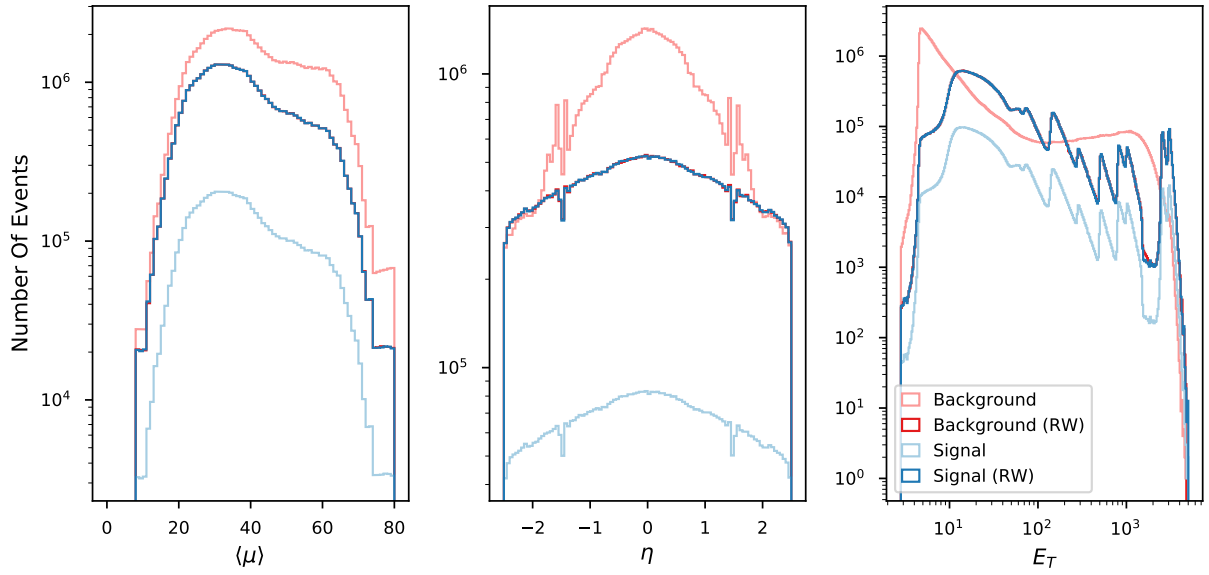


Figure 22: The result of the reweighting done for the MC photon dataset

Measures of Performance

The task at hand is one of binary classification. The end goal of an optimizer is to label as many objects given to it correctly, and ultimately the performance of a model is judged by this ability and this ability alone. However, given the complexity of the problem, the size of the phase space and all the quirks and nuances of working with a detector such as the ATLAS detector, one measure of performance on the whole evaluation set will not be sufficient. Therefore, a thorough analysis of the performance as a function of E_T, η, and $\langle\mu\rangle$ is vital to encompass all the details of the model. Nevertheless, the measure of performance in MC will be evaluated using a Receiver Operating Characteristic (ROC) curve.

The Receiver Operating Characteristics Curve In binary classification a model can produce four possible outcomes: the model can correctly label a signal object, a true positive (TP), or a background object, a true negative (TP), or it can wrongly label a signal object, a false negative (FN), or a background object, a false positive (FP). From here we can define the following two expressions, the true positive (TPR) rate and false positive rate (FPR):

$$TPR = \frac{TP}{TP + FN} \quad FPR = \frac{FP}{FP + TN}$$

Given TPR is fraction of total amount of signal in the sample labeled correctly by the model and FPR is the total amount of background in the sample labeled incorrectly by the model, it is convenient to rename them signal efficiency and background efficiency[18], respectively. Given the fixed cuts of the ATLAS working points these rates are constant for a given sample. Modern machine learning algorithm

[18] $\epsilon_{bkg} = (1 - BackgroundRejection)$

predict a score between 0 and 1, where a score approaching the extremities signals confidence in a signal label (1) or background label (0). However, at some point a threshold has to chosen, for which an object scoring above the threshold is classified as signal, after which the rates will be constant for the model as well.

The scores of two toy models are visualized in figure 23. Toy model 1 has the scores of signal and background objects shown in blue and red, respectively, while it is green and orange for toy model 2. It is clear from the predictions alone that toy model 2 will for certain thresholds have a favorable number of TPs and a lower number of FPs, leading to a higher TPR and a lower FPR. However, the extend of this is very well visualized by scanning from left to right in figure 23 and calculating the FPR and TPR along the way. The resulting curve with TPR on the x-axis and FPR on the y-axis is a ROC curve.

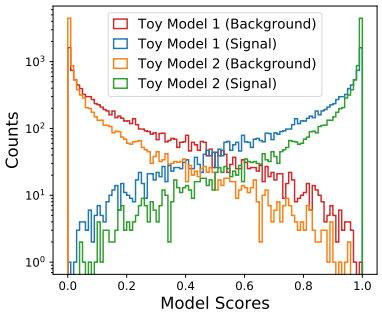


Figure 23: Score distributions of two toy models

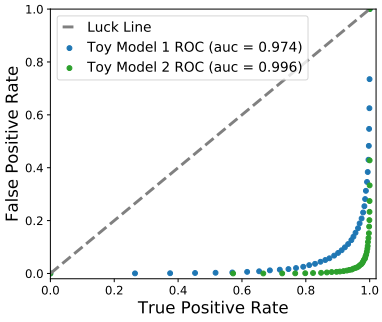


Figure 24: The ROC curve corresponding to the score distributions of the toy models

The result of scanning between 0 and 1 in increments of 0.025 on the scores of the two toy models is shown in figure 24. Here it is clear that toy model 2 allows for a low FPR for a significantly higher $TPRs$, exactly as expected from looking at the scores. A ROC curve is most often accompanied by a measure of Area Under Curve (AUC), which intuitively is a measure of the area under the curve. An AUC of 1, would be a perfect classifier. Note that here and consistently throughout this thesis, the number quoted as AUC is in fact area over curve ($1 - auc$). This is a result of the fact that ROC curves will be shown with FPR as a function of TPR, where an increasingly competent classifier approaches the bottom right corner, as opposed to the top left corner. The grey dotted line shown in figure 24 is the ROC curve expected from simple guesswork, meaning that the scores of the classifier overlap completely. It is therefore clear that the ROC curve and the corresponding AUC is a measure of separation between the score distribution of the background and signal objects. Given that the score distribution of a given classifier reflects an ability to separate signal and background in a high dimensional parameter space, we are directly probing the separation of signal and background objects. AUC allows for a one number representation of the ROC curve and for the nicely behaved ROC curves of the toy models, this number very well reflects the relative performance of the two models.

However, given the fixed rates of the working points, such a number does not exist. Therefore, in order compare our classifiers with ATLAS working points one has to define a measure of improvement, in this case an improvement factor, given by the following.

$$Imp = \frac{FPR_{ATLAS}}{FPR_{ML}} - 1 = \frac{FP_{ATLAS}}{FP_{ML}} - 1$$

Where FP_{ATLAS} and FP_{ML} refers to the number of false positives of the ATLAS working point and the classifier being compared, respectively. This comparison only truly makes sense if one matches the TPR of the classifier with the one of the working points. A similar improvement factor can be calculated by matching the FPR and

comparing the number of true positives. This can also be used to compare two classifiers, by matching a given TPR or FPR. Matching a given rate is achieved by applying a threshold to the score distribution, where the threshold is chosen for the rates to match.

Having explained the methods of evaluation, the specific machine learning algorithms used in this thesis will be described, in the following sections.

Random Forest (RF)

At the heart of the random forest lies the decision tree, possibly the simplest idea of a machine learning algorithm. A decision tree is a long nested if-else statement, wherein the conditions are cuts on variables[19].

[19] One can also propagate through a decision tree using probabilities

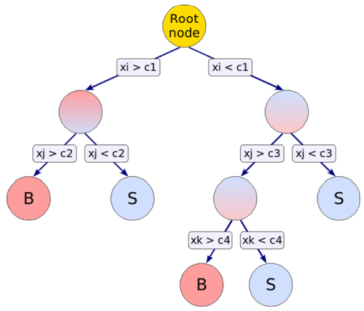


Figure 25: A decision tree, with the root node containing all data. A given cut c_1 is applied on a given variable x_i. The data left in each of the nodes then have another cut c_2 or c_3 applied on a given variable x_j, allowing for further separation. Finally on one node a last cut c_4 is applied on a given variable x_k. Figure taken from **[hoecker_tmva_2007]**

For a problem such as the one tackled in this work, a decision tree could be created following the recipe in figure 25. One would design the cuts as to perform as well as desired on the training set, test for over-training on a validation set and when satisfaction has been reached the tree would be evaluated on previously unseen data. This involves tuning a large amount of cuts and as the number of variables increase, this whole framework quickly becomes complicated. The next step is to automatize the cutting. This is done by introduction of a metric, a value to be optimized in order to select the best cut on a given variable. The metrics commonly used are gini impurity, information gain, and entropy. This allows the design on a decision tree that given enough nodes (a deep enough tree) reaches the desired purity. It is clear to realize that a perfect decision tree has enough nodes in the final layer to encompass every single datapoint. This, however, provides no generalizability, which is perhaps the most important feature of a classifier.

The random forest deals with this by creating several trees and applying the concept of bagging. Each tree will be given a fraction of the features (variables) and a fraction of the dataset. This, along with, proper tuning of the depth allowed, splitting criteria, and more, allows one to control the building of trees with the use of a validation set and only continue to build trees, while the model performance is generalizing. The evaluating metric used for the forest based models were AUC, meaning that trees were built as long as the AUC on the validation set increased.

Boosted Decision Trees (BDT)

The BDT is another decision tree based classifier. It creates a forest of trees and take advantage of the information gained by the previously built trees, when building a new one. This is done via a method called boosting. Effectively, boosting applied a higher weight to objects that were previously labeled incorrectly. This means that as the forest grow the incorrectly labeled objects will impact the splitting criteria more and more, forcing the trees to find splits that correctly

label these events. The method of boosting used for the BDTs in this work is gradient boosting.

The Light GBM Framework

LightGBM (LGBM) is an implementation of tree based algorithms, with various different methods of boosting, including RF (no boost) and gbdt. The largest advantage of LGBM is the optimization of speed. The growth of the tree is evaluated leaf-based as opposed to a classic level-wise growth, this along with highly optimized algorithms to select histograms and apply cuts.

- *Learning rate (LR)*. A minimal step size while progressing in the loss. The step size varies for different boosting types but the LR always acts as a minimum.
- *Number of leaves*. The total number of leaves a tree can contain. This parameter can be used to deal with overfitting.
- *Max depth*. The maximum depth, in layers, of the trees. This parameter can be used to deal with overfitting.
- *Minimum data in leaf*. The minimum amount of data in a leaf. A split will not be made if it results in a leaf with less than this value. This parameter can be used to deal with overfitting.
- *Minimum sum hessian in leaf*. The minimum sum of the hessian in one leaf. If a split results in a leaf with a sum hessian less than this value, it will not be made. This parameter can be used to deal with overfitting.
- *Bagging Fraction*. The fraction of data chosen to build the next tree. This reduces the strength of single trees while giving the full forest a larger ability to encompass the whole of phase space.
- *Bagging frequency*. The frequency of bagging.
- *Feature Fraction*. Like bagging, this selects a fraction of the features for building trees.
- *Minimum gain to split*. The minimal gain in gini impurity to perform a split.

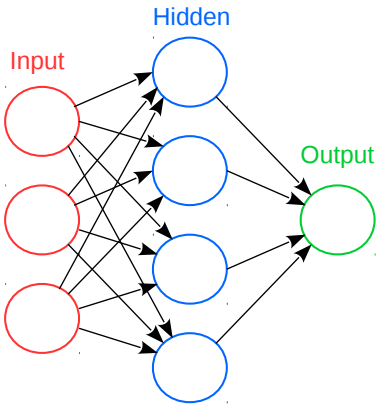


Figure 26: The overview of a neural network consistent of fully connected layers. The first layer is always an input layer, the size of which depends on the input dimensions, after which follows as many hidden layers as desired, the output of which is connected to a single output neuron.

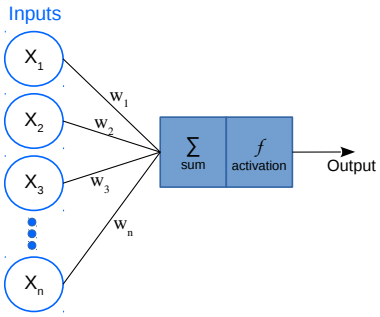


Figure 27: The artificial neuron takes n inputs of value $x_1 ... x_n$ and creates a linear combination of the inputs, accounting for their weights, $w_1 ... w_n$. This is then brought through a non-linear activation functions

Neural Networks

The universal function approximaters, also known as neural networks[**goodfellow_deep_2016-1**], have to no surprise become the haut monde in the world of machine learning. They are extremely versatile and multiple expansions on the base concept has led to neural networks being a cornerstone in solving many incredibly complicated and interesting problems, such as computer vision, speech recognition, and computers playing video and board games. The neural networks trained in this

Name	Expression
ReLU	$f(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } x \geq 0 \end{cases}$
LeakyReLU	$f(x) = \begin{cases} x\alpha & \text{if } x < 0 \\ x & \text{if } x \geq 0 \end{cases}$
Swish	$f(x) = x \cdot \frac{1}{1+exp(-\beta x)}$

Table 13: The activation functions considered for this work, where α and β are trainable parameters.

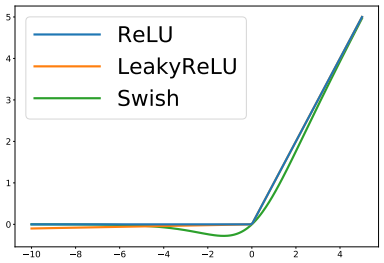


Figure 28: Activation functions tested during this work.

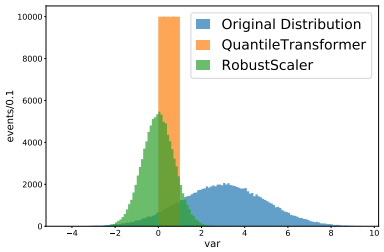


Figure 29: Comparison between QuantileTransformer and RobustScaler on a toy variable.

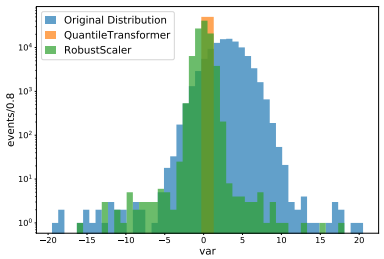


Figure 30: Comparison between QuantileTransformer and RobustScaler on a toy variable, with outliers.

thesis are Feedforward Neural Networks (FNNs) (figure 26), meaning the information only moves in one direction, from input to output. At the core of these networks are artificial neurons, an activation function, back-propagation, and a loss function. An example of the neurons in figure 26 is shown in figure 27. The reason for the non linearity of the activation function is that this allows the network to learn arbitrarily curved functions, rather than straight lines of the strictly linear interpretation. The learning is done by changing the weights of the connections in the network. The changing is done after each batch, which is an important hyperparameter of the network. The magnitude of this change is dependent on the optimizer. Common for all optimizers is that the gradients used are calculated using backpropagation, this limits this complexity of the activation functions. The optimizer used for this work is the Adam optimizer[**kingma2014adam**]. Several activation functions have been considered for this work and will be described in the next section.

Activation functions Many different activation functions have been considered for neural networks throughout time. It is still an active field of research and the landscape is ever changing. For this work three activation functions have been considered: Rectified Linear Unit (ReLU)[**maas_rectier_nodate**], Leaky Rectified Linear Unit (LeakyReLU), and Swish[**DBLP:journals/corr/abs-1710-05941**]. The first two are stable tools in any deep learning toolbox. Swish is a newer addition and promises improvements over exactly ReLU and LeakyReLU.

Optimization and useful tools

Preprocessing A common step before training an algorithm is different steps of preprocessing. This includes many different methods, one of which deals with the scale of the input features. A lot of algorithms perform significantly worse on input features with different scales, therefore, a common step is to scale the distributions to a mean of zero and a variance of one. In python, this can be done by applying the RobustScaler from SK-learn[**scikit-learn**]. However, the RobustScaler does not deal well with outliers and keeps long tails. An example of the RobustScaler working nicely has been visualized by applying the scaler on a single Gaussian distribution ($N_{samples} = 100000, \mu = 3, \sigma = 3$) (figure 29). If one simulates outliers by adding 100 random samples from a ten times wider Gaussian distribution ($N_{samples} = 100, \mu = 3, \sigma = 30$), the results (30) clearly show the tendency of not scaling outliers properly and given many variables with long tails this would be an insufficient scaling. Also shown in figures 29 and 30 is the output of another preprocessing tool, the QuantileTransformer[**scikit-learn**], which transforms the whole distribution to a uniform distribution in range $[0, 1]$. It reduces the impact of (marginal) outliers and has a tendency to spread out the most frequent values. This transformation is non-

linear, which makes variables measured on different scales more easily comparable, while potentially distorting linear correlations between variables on the same scale. This method proved extremely potent and for all neural networks trained in this thesis, the QuantileTransformer is the only preprocessing step applied before training.

*Scikit Optimize (skopt)[***scikit-learn***]* is a framework used extensively in this work to optimize the hyperparameters of the models. Specifically, *gp_minimize* and *rf_minimize*. The strength of these modules is their versatility, ease of use and performance. The strongest of the two is the *gp_minimize*, which utilizes Gaussian processes to map the complex parameter space by sampling and reach convergence. The challenge for tree-based models is that the hyperparameter space is very flat, performance of any reasonable set of hyperparameters perform very similarly and generally the main attribute to access is the generalizability of the model. This affects the performance of *gp_minimize* by decreasing the ease of use. It does not work right out of box and the noise term has to be tuned in order to promote exploration. If this is not done, the module converges to a corner of phase space more often than not. As an alternative, *rf_minimize*, has been deployed late in the work to ensure convergence, which given the shape of the hyperparameter space, should work to a reasonable degree. One could optimistically gain a few percent in performance but a choice has been made to use *rf_minimize* given time constraints and the amount of models trained.

Learning Rate Finder (LRF) The learning rate (LR) is a vitally important hyperparameter of most ML algorithms. Properly tuning this variable can greatly optimize your training time, which for some algorithms is very expensive. To assist with this, two frameworks have been applied: one for the BDTs and one for the networks. The LGBM LRF works simply by running a training with base parameters while applying different learning rates, both stable and decaying ones. In essence, one can code up any schedule imaginable as a function of the number of trees built. The goal is to achieve best and smoothest convergence and considering that training time is not very expensive, it is reasonable to achieve this goal without considering speed. An example of the output from the LRF can be found in the appendix

For the networks, on the other hand, training time is very expensive. This can be dealt with in many different ways, one can apply different types of learning rate schedules: one-cycle, cyclical, to name a few. The networks shown in this thesis use a cyclical learning rate (CLR), which provides a triangular learning rate schedule, that steadily increases from a minimum LR up to a maximum LR and back down, across each batch. These learning rates are optimized using a framework that scans a wide range of learning rates in a couple of batches, plotting the loss as a function of LR, allowing one to select the optimal range for learning. The cyclical learning

rate is provided as a callback for keras[**chollet2015keras**], which is used along with a Tensorflow[**tensorflow2015-whitepaper**] back-end for the neural networks trained for this work.

*SHapely Additive exPLanations (SHAP)[***lundberg_consistent_2018***]* The question to answer is; "How does one consistently and accurately measure the impact of features on the output of an algorithm?" The answer has been found to be SHAP and has been used extensively in this thesis to study performance and behavior of the algorithms, to ultimately decide on which variables to include and provide important insight into the inner workings of our models. The inner workings of SHAP is very complicated and use linear regression, game theory and more to estimate feature important.

Particle Identification

The following chapter will walk through all the training done on both photons and electrons in MC and data. The results of each training will be evaluated as they training of the model is introduced. The goal of this is to show the work process as it unfolded and explain the important details and findings along the way.

Training in Monte Carlo

As mentioned in chapter , the MC dataset is high statistics and contains both converted and unconverted photons as well as a background distribution containing background photons, electrons and a few hadrons. Traditionally, converted and unconverted photons are dealt with separately and as mentioned in chapter the cuts applied by ATLAS are optimized in bins of η and E_T. Following this recipe, one could train a model in each bin as well as with both unconverted and converted signal and background types. This could ultimately lead to a strong versatile framework, however, it would require an immense amount of training time as well as statistics that challenge even the dataset created for this thesis. A choice was made to attempt an inclusive training: only one model was trained, one that covers the whole η, E_T phasespace and all types of signal and background. This is a heavy task to ask of one model, however, the exorbitant amount of time gained from not having to train all these individual models should provide plenty of time to make an inclusive training work.

The evaluation of such a model still requires separation into converted and unconverted particles and a binned measurement of signal and background efficiencies. Not only are these logical checks to make but are also vital for the possibility of comparison with the already established working points, provided by ATLAS. The ultimate goal of evaluation is to show an as reasonable comparison with the Loose and Tight ATLAS working points as possible, within the reaches of my dataset. Since the working points are optimized in bins, showing an overall performance does not necessarily provide misrepresentation of their performance, however, since the aim is to show an improvement, it is important to know that this improvement is not localised in phasespace.

A long process predates the final variable sets used for the MC training on photons. A vital component in this selection of variables in the SHAP framework as descriped in chapter . As mentioned be-

fore an abundant amount of information is available about a given photon and much of this information has been scrutinized for potential involvement. A preliminary list of 123 variables was chosen for further examination and through a couple of iterations of shortening the list and retraining, it became clear that including more than just a few extra variables had little to no impact on performance. A finalized list of variable sets were decided upon. Each expanded set includes the older sets as subsets. The variable sets are described in

Table 14: The datasets used for photon training, each set contains the previous sets as a subset. Models trained on the datasets will be represented by the color (or a similar shade) listed in the table.

Name	Nvars	Color	Vars
Showershape (pdf)	11	**Blue**	Rhad1 Rhad Reta weta2 Rphi Reta fracs1 DeltaE Eratio f1
Conversion (conv)	15	**Orange**	ConvertionType ConversionRadius VertexConvEtOverPt VertexConvPtRatio
Binning (bin)	18	**Green**	η E_T $\langle\mu\rangle$
Extra (ext)	22	**Red**	*maxEcell_time* *maxEcell_energy* core57cellsEnergyCorrection r33over37allcalo

table 14, where the name briefly describes the contents of the variable sets, and the shortened name in parentheses will be the reference of the specific variable set throughout the evaluation. The color of the sets will be the color of the LGBM MC models trained on the respective variable sets. Other models trained on the same set of variables, will be given a different shade of the same color. The SHAP values of a model trained on the extra variable set is shown in figure 31. This complete list is representative of a given variable's performance within its own set to a reasonable degree, however, small shifts are expected. The list shows an overall strong performance of the showershape variables along with clear evidence for the inclusion of the extra variable set. Another interesting note is the performance of the transverse energy of the photon, that has no discrimination power, due to the reweighting, described in chapter . The theory is that it is being used as a high level variable in the models, which allows for determining regions of phase space. In effect, the models are creating their own bins, which in turn allows them to better deal with the

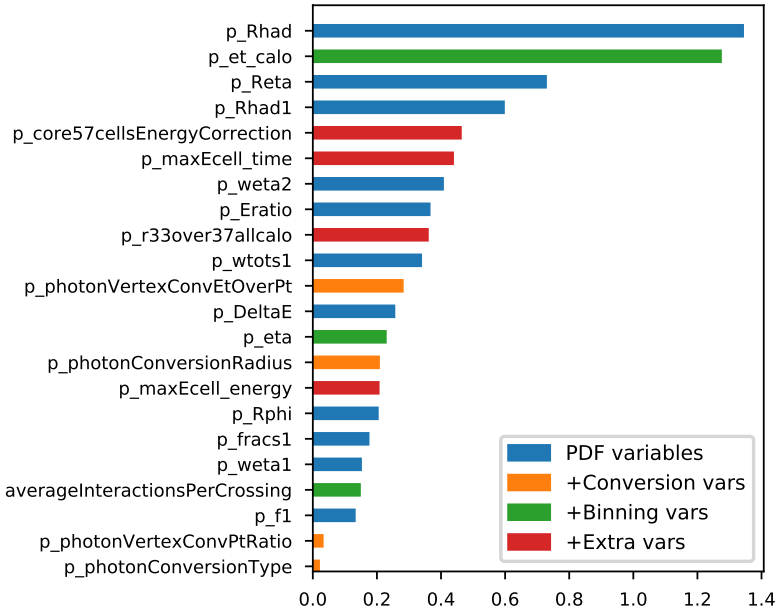


Figure 31: The SHAP value ranking of a model trained on the extra dataset.

slightly different distributions of signal and background present in these bins. The same can be said for η, where one would also expect a learned binning to allow the model to better deal with changing showershapes. Note that a variable with no discrimination power will almost always have an extremely low score (< 0.0001), which is why this suspicion was originally raised. This is exactly the behaviour one could hope for by including the binning variables and it is therefore a welcomed effect.

To allow for continuity the trained models will be named according to a naming scheme, as to clearly inclose the most important information regarding a given model. The naming scheme is the following:

$$Modelname : "MODELTYPE"("DATATYPE","varlist") \quad (7)$$

Such that a MC LGBM model trained on the showshape variables will be called $LGBM(MC, pdf)$. Models trained on the variable sets from table 14 have their overall performance shown in figure 32 for the unconverted and converted channels independently. As described in the ROC curve is a visualization of ones ability to select signal (signal efficiency) and background (background efficiency). The ATLAS working points are shown along the models and a clear overall improvement in performance is demonstated by these figures. The improvement is clearly largest for the converted channel, which is to be expected, given the power of the conversion variable set on this channel. An interesting behaviour that becomes apparent from the separation into channels is the obvious differences in variable importance. $LGBM(MC, conv)$ only improves in the converted channel and seemingly manages to perform as well as $LGBM(MC, pdf)$ on the unconverted channel. This is a comforting behaviour, since it is

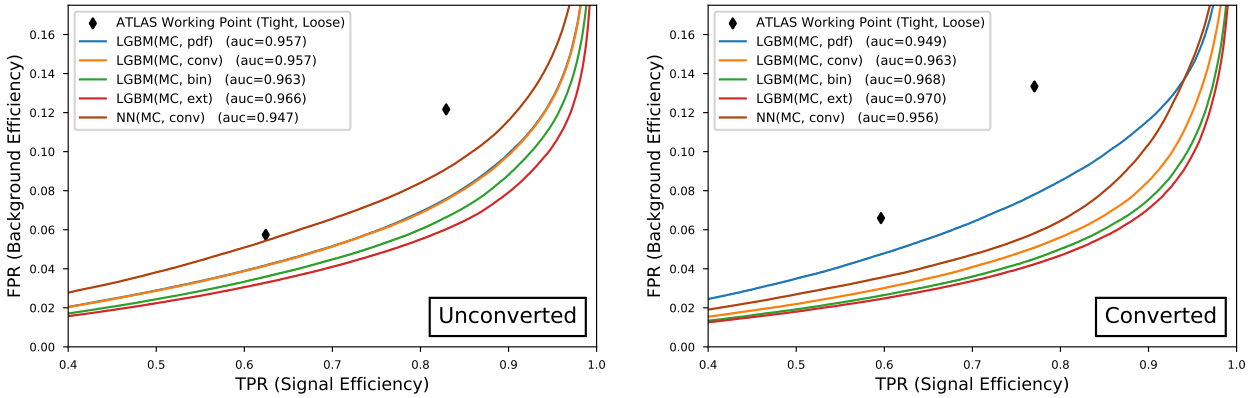


Figure 32: The roc curves of the photon models in the unconverted (left) and converted (right) channels. The ATLAS working points are shown in black.

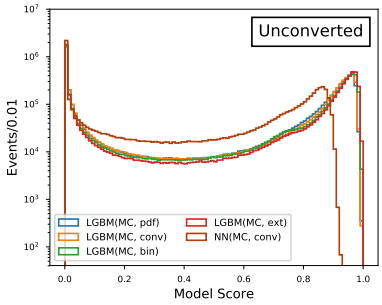


Figure 33: The score distribution of the photon models in the unconverted channel.

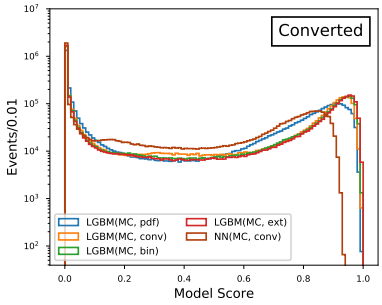


Figure 34: The score distribution of the photon models in the converted channel.

vital for an inclusive training to make sense. On a similar note, the extra variable set seems perform slightly better in the unconverted channel.

Figures 33 and 34, clearly shows that the $LGBM(MC, pdf)$ model struggles in the converted channel, whereas the more complex models converge nicely. Another thing that along with the ROC curves becomes very clear is the poor performance of $NN(MC, conv)$. it simply does not perform comparable to the LGBM models, the reason is most likely due to dataset, while the QuantileTransformer allows to proper training, the many quirks of the dataset seemingly affects the neural networks more. This confounded with the training being done without weights simply does not allow for a competitive comvergence. For now the performance of $NN(MC, conv)$ will be included in the evaluation plots, however, the focus will be upon the well performing models.

For an indepth look at the performance of the models, their performance as a function of $\langle\mu\rangle$, $|\eta|$, and E_T, are compared to the ATLAS working points. This is done by matching the signal efficiency of the models resulting in a comparable background efficiency. The matching is done by finding a threshold that results in an equal amount of true positives. These performance plots are shown on pages 67 and 68, where the models are compared with the loose and tight working points respectively. These plots all supports the overall performance promised in figure 32, with all LGBM models comparing favourably to the working points, as well as model complexity leading to better performance holistically. This statement holds true for bins in the $\langle\mu\rangle$, η, and E_T phase space, across both working points. Similarly, the improvement is best in the converted channel, while not as clearly as in figure 32. The note on the performance of the $LGBM(MC, conv)$ model in the unconverted channel is further strengthened with it only performing slightly worse than $LGBM(MC, pdf)$ in a single bin ($E_T > 250GeV$). An encouraging sign is that in almost every single bin the trend in performance is shared between the working points

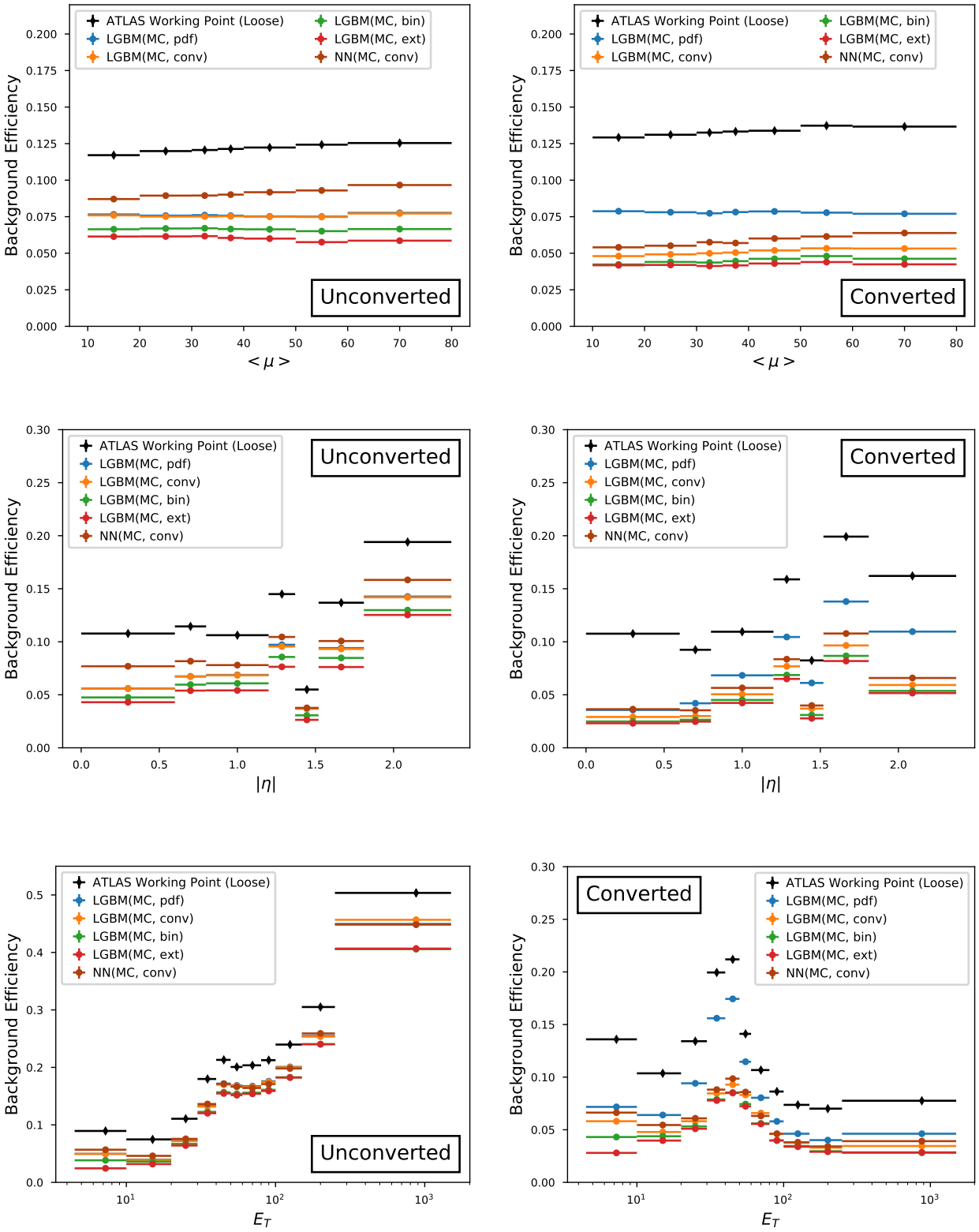


Figure 35: The background efficiency of the **loose** working point and the photon models as a function of $\langle\mu\rangle$, $|\eta|$, and E_T in the unconverted (left) and converted (right) channels. The signal efficiency is in each bin matched to the one of the working point.

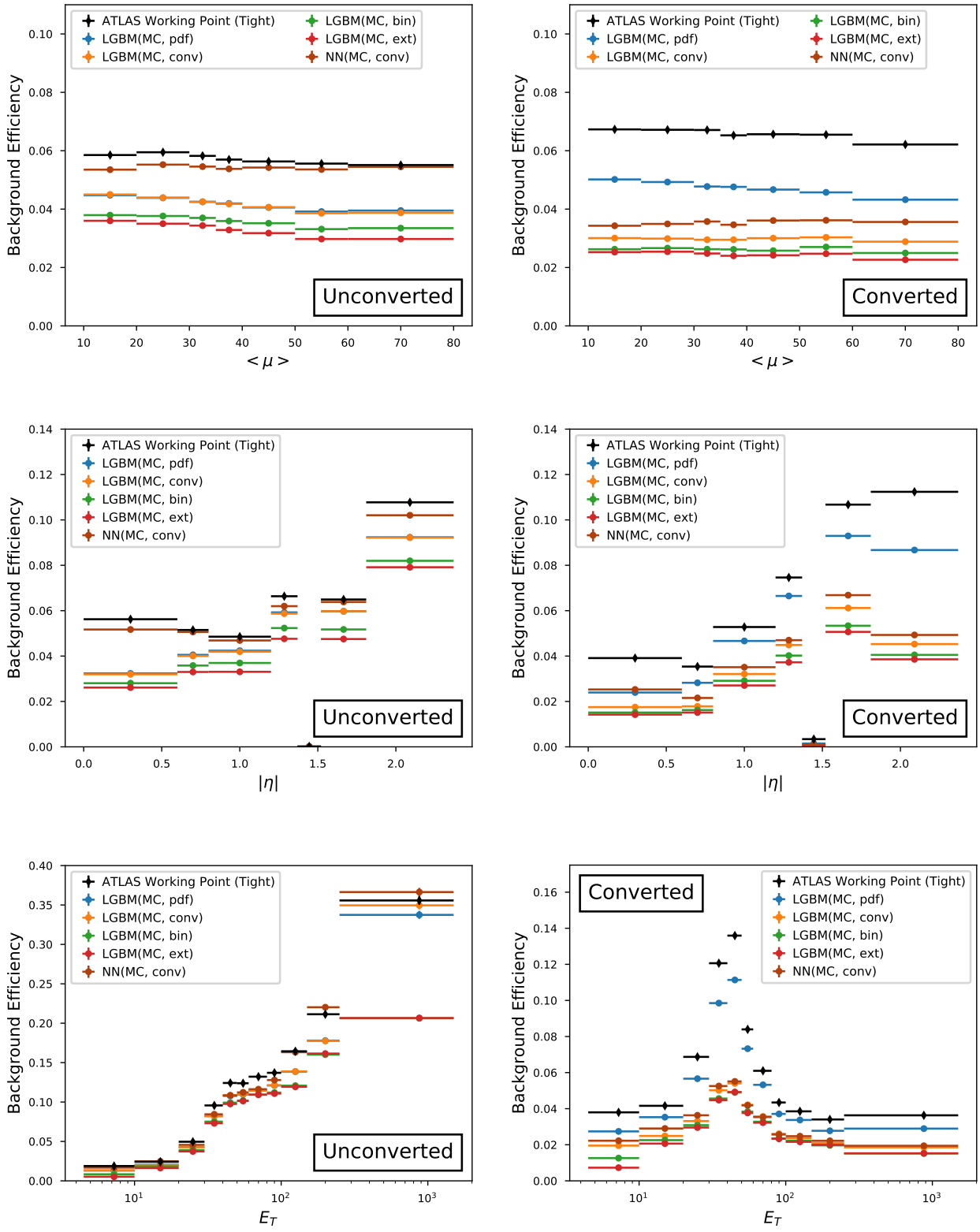


Figure 36: The background efficiency of the **tight** working point and the photon models as a function of $\langle\mu\rangle$, $|\eta|$, and E_T in the unconverted (left) and converted (right) channels. The signal efficiency is in each bin matched to the one of the working point.

and the models. The most noteable differences are found in the converted channel ($|\eta| > 1.52$), where all models, but $LGBM(MC, pdf)$ deviate from the working points, and in the unconverted channel ($E_T > 250GeV$) in the comparison with the tight working point, where $LGBM(MC, ext)$ as the sole model diverges. All in all, there are no surprises and the improvements are well distributed across the tested binning and the models are behaving nicely.

However, some of the trends of the working points that the models mimic consistently are a little worrying. The seemingly increasing performance as a function of $\langle\mu\rangle$ is highly unlikely to be due to it actually being easier to select signal at higher pileup. It is very well known that particle identification is challenged by an increasingly noisy environment, of which $\langle\mu\rangle$ is a good measure. Similarly the performance as a function of E_T exhibits a large decrease in performance as a function of E_T in the unconverted channel and a peaky structure in the converted channel. Both of these trends are unexpected, particle identification is expected to become easier as energy increases[20]. However, the most apparent cause of worry is the fact that the crack ($1.37 < |\eta| < 1.52$) is seemingly the best part of the detector for both working poins and both models. This is incredibly unlikely to be true, as described in chapter , this is the part of the detector with the least active detector material. For all of these worries exists an explanation: the working points are optimized against a very specific background composition and can, therefore, behave quite sporadically on a vastly different background. This background sample used for this work is in fact quite different. An insight into this

[20] One might have heard an electron at high energy being refered to as a christmas tree

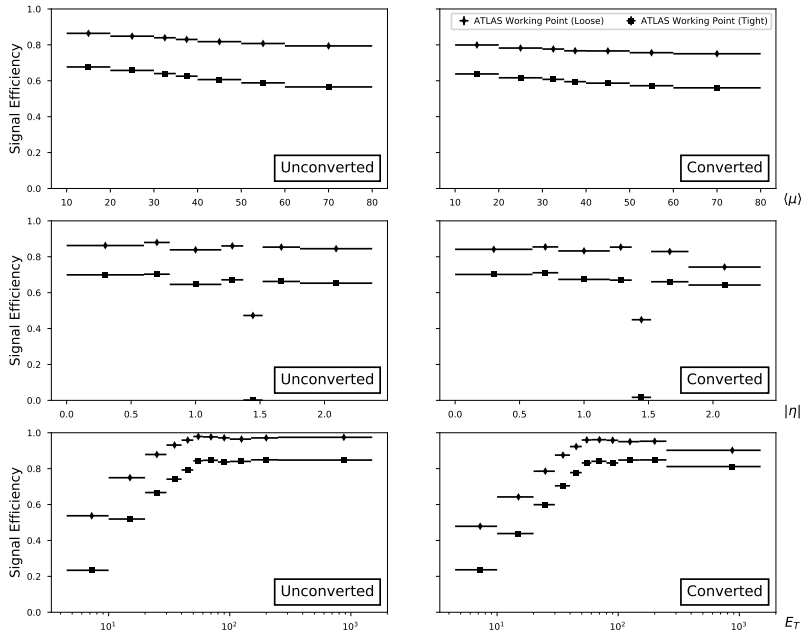


Figure 37: The signal efficiency as a function of $\langle\mu\rangle$, $|\eta|$, and E_T of the ATLAS working points on the photon MC dataset. Note that the tight working point is a subset of the loose, meaning it always selects less signal.

matter can be gained in figure 37, where the signal efficiency of the working points is shown. This figure clearly shows that the working points are selecting signal entirely as inteded. As a function of $\langle\mu\rangle$

and $|\eta|$ they select a slightly decreasing amount of signal, with the exception of the crack ($1.37 < |\eta| < 1.52$) wherein the loose working point has it's signal efficiency cut in half and the tight working point collapses completely, due to the exclusion of this area during training and optimization. As a function of E_T the signal efficiency starts low and rises to stable values. This behaviour is exactly the way the working points are designed to function. They deliberately select less signal in parts of phase space that would otherwise bring in too much background. This means that the sporadic behaviour seen in the background effiency plots has to be explained by background composition. The peaky structure in the converted channel could be explained by the fact that a large fraction of the background is in the bins, where the working points struggle to select a large fraction of the signal. This is simply due to the vastly different background composition from the one the working points are optimized against and is to no fault of the working points.

However, while the working points have the advantage of being well tested and stable on a background composition that very well imitates the circumstances in data, this is certainly not the case for the models. The model are trained on this seemingly problematic background composition and should therefore ideally be able to show a stable behaviour on it. In order to ensure this a the model performance is tested for a fixed signal efficiency of 92% and the background efficiency across phase space is once again determined. The results of such efforts can be seen on page 71. The models manage to show the expected, stable behaviour, in 5 out of 6 cases. The performance of a function of $\langle\mu\rangle$ is slowly decreasing, the performance as a function of $|\eta|$ is slowly decreasing with a spike in the crack, and the performance as a function of E_T is increasing. This last part only being true for the converted channel. This last instability, was also clearly present for the working points and must therefore be an artefact of the background composition. The main argument behind this reasoning is visualized in figure 39, where the predictions of $LGBM(MC, pdf)$ in the highest energy bin ($250GeV < E_T < 1500GeV$), is shown with the different background types shown on separate axes. The logarithmic y-axis complicates the comparison between axes but even so it is very clear that $> 99\%$ of the background in the unconverted channel is background photons and relative to the converted channel they are far more likely to be signal-like. This severely impacts the background efficiency at a fixed signal efficiency and perfectly explains the vast differences between the two channels as a function of E_T. However, the background composition is far different from what one would expect in a physics analysis. This was already clear when the background efficiency of the working points were so sporadic at reasonable signal efficiencies.

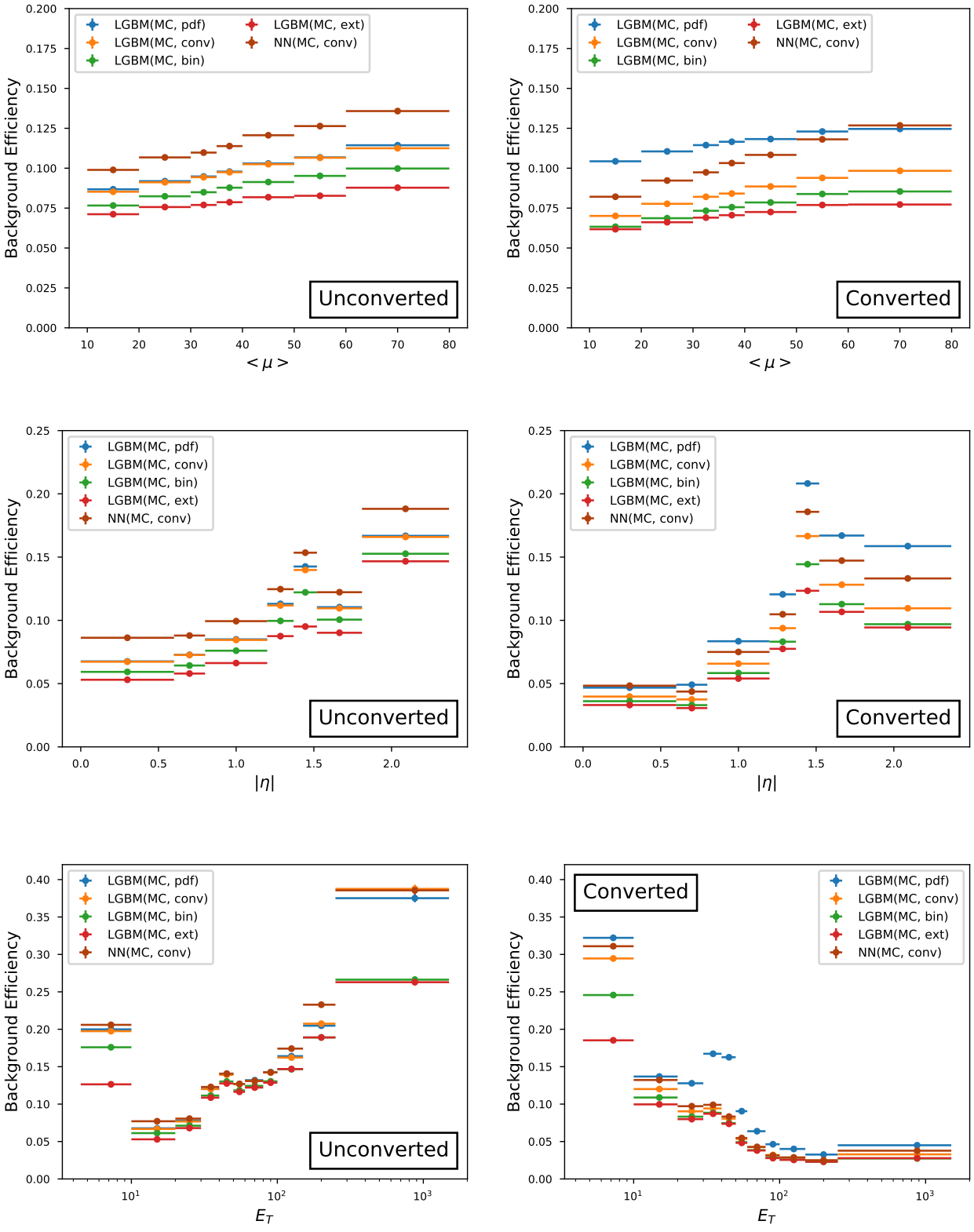


Figure 38: The background efficiency of the photon models in bins of $\langle\mu\rangle$, $|\eta|$, and E_T for a fixed signal efficiency of 90%.

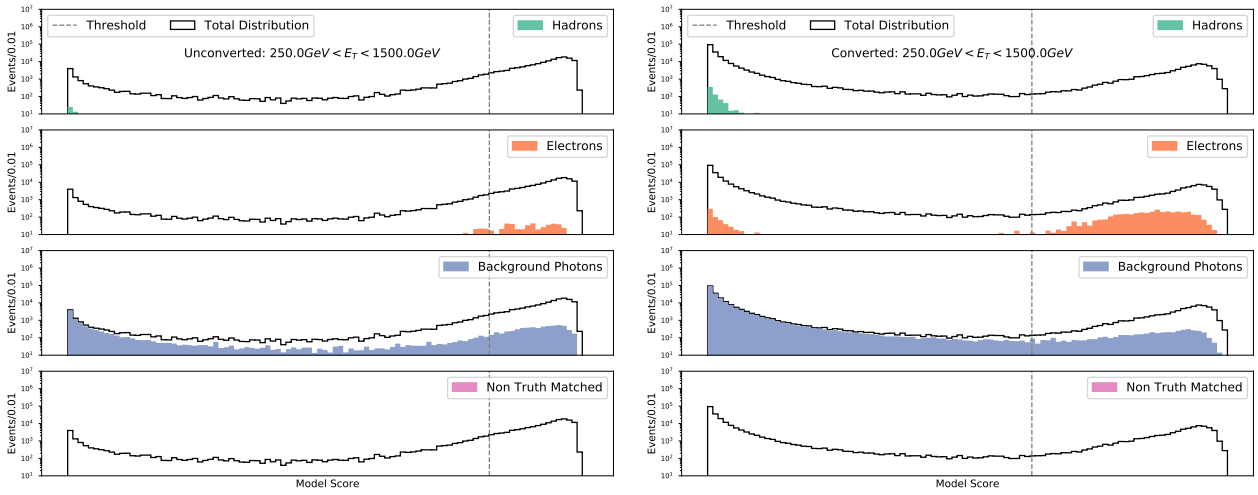


Figure 39: The score distibutions, in the high energy bin, of the different background types compared to the full distribution of scores. Note the log scale on the y-axis. It is clear to see the large difference in the distribution of background photons in the unconverted (left) and converted (right) channels.

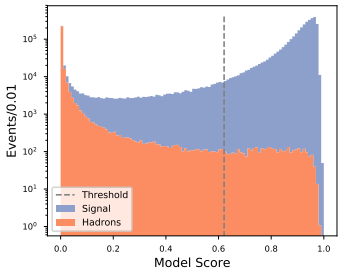


Figure 40: The score distribution of signal against hadronic background for the unconverted channel

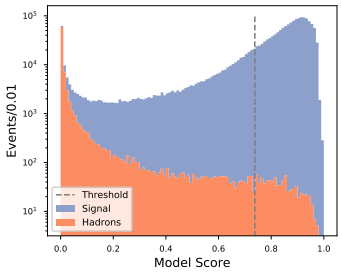


Figure 41: The score distribution of signal against hadronic background for the converted channel

The working points are mostly optimized against a background of hadrons and the models manage to effectively discriminate against this background type, visualized by the overall stacked scores of background hadrons and signal photons seen in figures 40 and 41, where the threshold represents a signal efficiency of 92%. While this is an overall distribution and does not necessarily reflect across all of phase space, the vast majority of hadrons are very well separated from signal by the weakest of the models. A distribution of hadrons at high E_T simply is not possible to show, given the lack of reasonable statistics. This exact problem led ATLAS to the use of a data-driven background sample. Note that the ones present in figures 40 and 41 are very well discriminated against. The conclusion is that one can trust the models to behave nicely as a function of E_T in data and ultimately the behaviour of the models are proven to be solid and stable, with a promise of large improvements. All of this being in MC.

Having established that the MC trained models behave nicely and manage this in spite of a rather unusual background sample, the next logical step would be to evaluate the performance on a resonance. The resonance chosen is $Z \to ll\gamma$. Here one could gauge the improvements of the inclusive training on a sample one would be able to select in data. However, this work was led in another direction due to the interest of the MC training quickly being replaced by a vision of training directly on data. The main motivation behind this vision is the very well known fact that improvements seen in MC might not translate to data at all. This is due to the both unfortunate and at the same time wonderful problem: the real world is extremely complex and while the MC simulations deployed by ATLAS are absolutely state of the art, despite decades of work, they are simply not

perfect.

Perhaps, despite the simple fact that data will always be messier than MC, despite the fact that we won't with 100% certainty know if our labels are correct and despite a severe hit in statistics, the models trained in data will overcome all of these obstacles and perform better, because of the simple fact that it has seen actual photons and actual background.

Label Confusion

No matter how one tackles the problem of training machine learning models[21] one will at some point have to guide the algorithm by telling it whether it is predicting correctly or not, which will ultimately shape its ability to learn to distinguish signal and background. This involves a decision, which in data will never be 100% correct that will lead to imperfect labels and this concept is referred to as label confusion, illustrated by figure 42. The presence of signal in the background label is referred to as signal confusion, whereas the opposite is referred to as background confusion. The naming convention is rather arbitrary but is chosen based on the impact achieved by simulating a given type of label confusion. Label confusion is unavoidable in most selections one would apply in data. A clear example of this the output of the electron selection, outlined in chapter . The invariant mass distribution shows a clear signal peak on top of an falling background. Signaling a presence of background in the signal label. One can reduce such a background by cutting tightly around the signal peak and/or applying an isolation requirement, however, this will lead to a significant loss of signal statistics and potentially bias the probes. Instead, one usually deals with this by fitting and subtracting the background, however, this only works on distributions and does not allow for determining whether a given object is signal or background. The conclusion is that any realistic selection in data that would allow for high statistics and unbiased

[21] Here I refer to supervised training

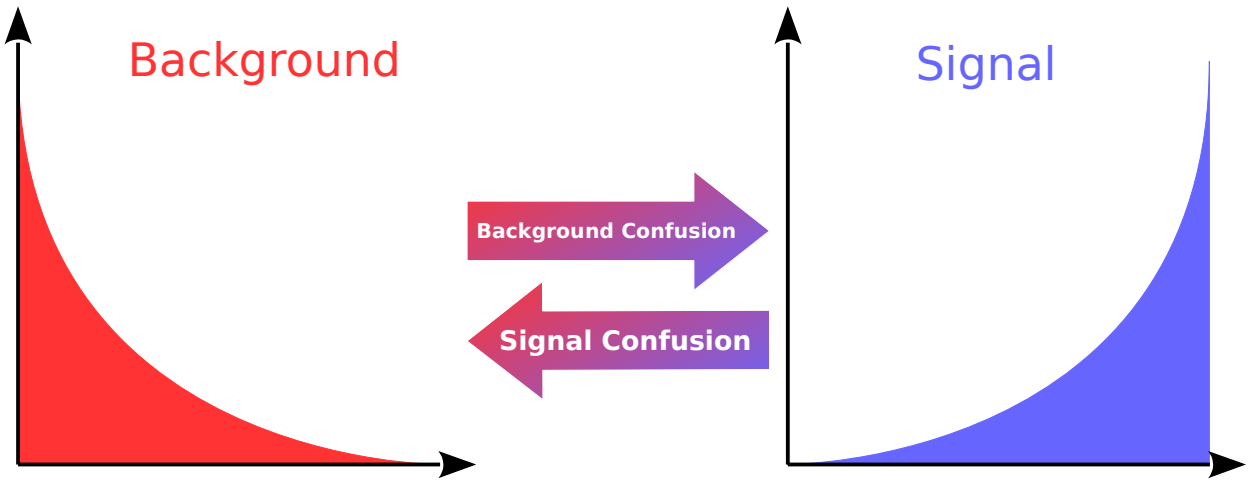


Figure 42: A visualization of the two types of label confusion tested. As indicated by the arrows, signal confusion is the concept of signal objects being labeled as background, while background confusion is the concept of background objects being labeled as signal.

probes will bring with it a degree of label confusion and if one hopes to train in data, the algorithms used will have to be robust to this.

This will be tested by simulating label confusion in MC. By exposing the LGBM implementation of both a BDT and RF to an increasing degree of both signal and background confusion. The goal is to test the algorithms to their breaking point, which involve confusions that are far beyond what is reasonable to assume is present in data.

Models will be trained and validated on an increasingly confused training and validation set and will afterwards be evaluated on an unconfused test set. This allows a gauge of the models ability to remain able to select the true labels, despite the confused training. The models will be exposed to a grid of signal and background confusion ranging from 0.0% to 60.0% in steps of 10%. Ultimately, the robustness of a model will be gauged on the relative degradation in performance with respect to the unconfused training. The switching of the labels is applied in a completely random fashion, giving a confusion that is entirely uniform within the signal and background distributions. This type of confusion is completely irreducible and the question becomes at what saturation of the respective distributions does the model lose the ability to distinguish the true distributions. The analysis will be done with signal photons against background hadrons, selected randomly from the MC photon datasets, to allow for a good initial separation with little signal-like background. This training (validation) set of 1000000 (100000) random signal photons and background hadrons has been chosen. The equal amount of signal and background events is chosen in order to equalize the impact of the two types of confusion. The evaluation is done on all the signal photons and background hadrons available in the test set. A LGBM RF and BDT were trained in each of the 49 bins spanned by the increasing signal and background confusion. On a technical note, the models were trained for 1000 trees with the training, validation and evaluation metric saved for each training step. This was done to make it possible to analyse the training afterwards. Early stopping was then simulated by selecting the predictions of the model with the lowest score on the validation set. Each model was evaluated with auc on the validation and evaluation sets and the resulting grid can be seen in figure 43. The top row represents the evaluation on the validation set, while the bottom row represents the evaluation on the evaluation set. This figure is incredibly hard to unpack, as it provides an untold amount of interesting results. Firstly, the robustness of the models displayed in this analysis is astounding, with very little relative degradation happening until extreme levels of confusion. Secondly, and perhaps even more interesting, when the only confusion present in the sample is background confusion (no signal in background, background in signal), the training is aside from a slight hit in overall performance almost entirely undisturbed, with practically no loss in performance on the evaluation set. This is perhaps the best piece of information one could get if one hopes to train in data. In many cases producing a selection, such that there is a nel-

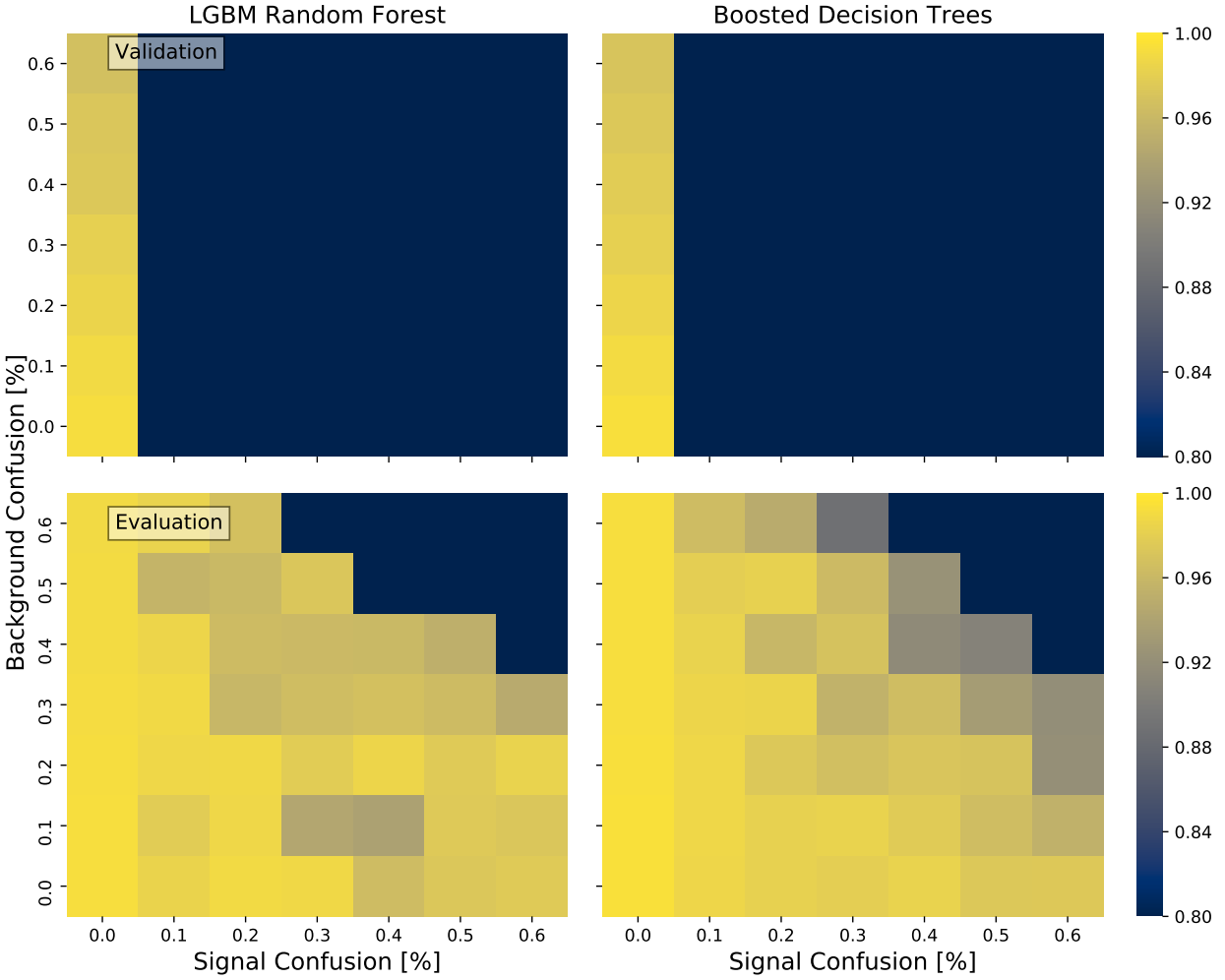


Figure 43: The (auc) evaluation of RF and BDT models trained with increasing signal and background confusion as labeled by the axis. The evaluation is shown for the confused validation and unconfused evaluation sets separately. The models manage to learn the true labels with only slight degradation in performance for incredible amounts of label confusion. While validation only managed to stay useful while the signal confusion is low.

igible amount of signal in the background is easily doable, especially if one is allowed to include a significant percentage of background in the signal label. On this statement alone, training in data should be deemed a possible endeavour. However, there are much more interesting information to be gained from this label confusion study. The first of which is the evolution of the model score distributions as confusion increases, visualized in figure 44 on page 77, where the scores of the BDT model on the evaluation set is shown, colored by their true labels. The unconfused training is shown in the plot nearest origo and if one follows the evolution directly right or up, one sees why the confusion types have been named as they have. The models simply lose confidence in the confused label, while still being able to separate the true labels. Naturally, confusion in both labels force the distributions closer. The gradient degradation of performance becomes very apparent if one applies the knowledge gained about score distributions, roc-curves and how this translates into performance. Total degradation of the models does not fully happen until confusions of the level 50/50, 40/60, or 60/40. A breakdown at these levels of confusion is expected, given the labels are practically equal signal and background. Beyond these levels of confusion one see models working exactly as intended. However, they are now effectively classifying true background as signal and vice-versa. Essentially a systematic breaking of our models has revealed very few flaws and their behaviour is as predictable as it is consistent. While this analysis has already gone way beyond the scope of the initial motivation it is important to state the fact that the RF simply behaves much better in the face of confusion, while the BDT is saved by early stopping. This is visualized nicely in figure 45, where it is clear that the validation set is of little help for both models. However, where the BDT manages to learn something that translates to the evaluation set in the first few trees it quickly begins to overtrain. This is exactly what the BDT is designed to do and turns out to be its biggest weakness, when faced with label confusion. The RF training curves fall out eerily similar across every single of the 49 bins, it does not overtrain once and is able to translate all of its learning on the training set directly onto the true labels. This makes the RF incredibly robust to label confusion. It also entirely removes the need for a validation set. One simply has to train the RF for a reasonable amount of trees matching the complexity of the problem and then one can be very certain to end up with an highly performing model.

Nevertheless, given the performance as a function of background confusion, where both models managed great training and validation curves. The better overall performance of the BDT in these bins makes the BDT more than eligible for training in data. It is also clear that applying early stopping in data should work as intended. All in all the currently established framework for training in MC will be the one used for training in data.

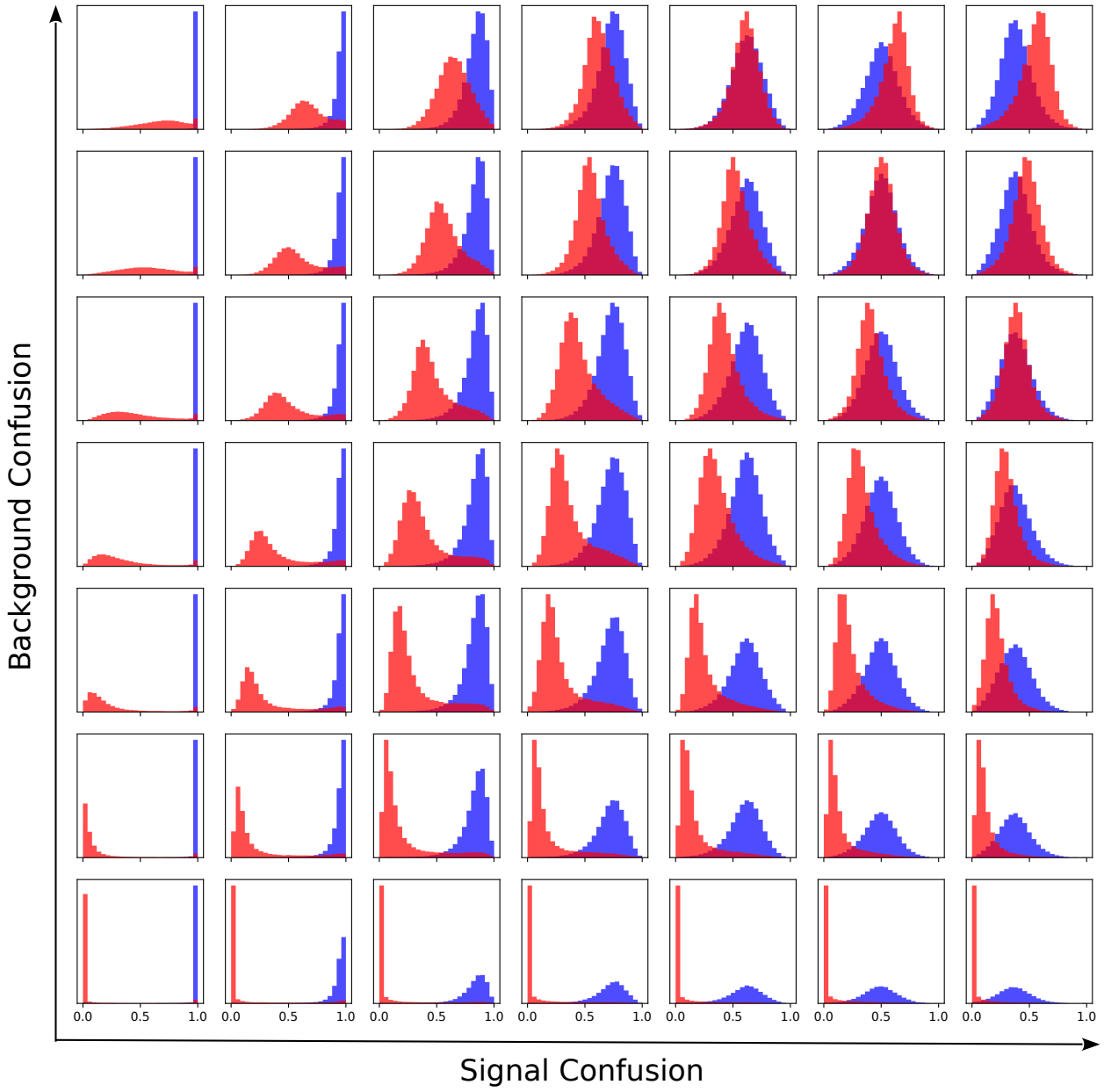


Figure 44: The score distributions of the BDT models on the evaluation set shown with increasing signal and background confusion. The plot nearest origo is the unconfused training, with signal and background confusion increasing in steps of 10% going right and up respectively. The signal and background distributions are each normalized to an area of 1

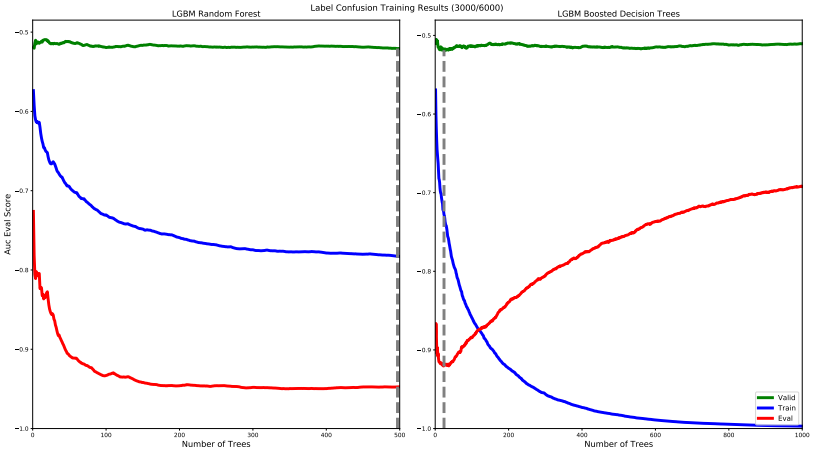


Figure 45: The running evaluation on the training, validation and evaluation dataset for the RF and BDT model trained on 30% signal and 60% background confusion. The RF shows no sign of overtraining and learning on the training set translates very well to the true labels, while the BDT suffers from massive overtraining, but is saved by early stopping.

Training in Data

Having established that our algorithms are robust to label confusion, to a reasonable degree of confusion, training in data should be possible with the methods already established for this work. However, the new challenge is to select labels in data, given that the luxury of truth matching is no longer a possibility. There are many ways of achieving labels in data and the goal of this chapter is to acquire labels that are as clean (unconfused) and unbiased as possible, as this should hopefully lead to the best training possible. The datasets available are selected using the *T&P* and *TT&P* selections as described in chapter , with the electron dataset being by far the least confused sample. The strongest tools available to select signal in these datasets are cuts in invariant mass and isolation, both of which has a possibility of introducing biases. Especially isolation, which is correlated with particle identification (highly so for background). This means that the potential information gained by training on the labels selected using an isolation requirement will be limited. However if one obtains an isolation measurement that is decorrelated from Particle Identification (PID), a cut can be applied on this measurement in order to select clean labels without further bias. This would be achieved by training a machine learning model, selecting isolated objects in the resonance, on isolation variables. The method of decorrelation will be described in the following paragraph.

Decorrelation of ISO and PID scores The goal is to acquire an isolation score that is (linearly) uncorrelated with PID. The reason for this is that such a variable will potentially be very powerful in selecting less confused and biased labels in data. This will lead to a better training and hopefully result in further gain when evaluating in data.

Normally when decorrelating one applies a rotation, in the two-dimensional space spanning the two variables. However, this changes the values of both variables, which is unsatisfactory. Therefore we have to apply a more complex decorrelation scheme, which is de-

scribed below. We define two variables:

$$x = \frac{pid}{\sigma_{\text{pid}}} \quad and \quad y = \frac{iso}{\sigma_{\text{iso}}}$$

Where pid and iso refers to the particle identification and isolation predictions, and σ_{pid} and σ_{iso} refers to the uncertainties of the distribution of predictions. Then the following is true:

$$\sigma^2(x) = \sigma^2(y) = 1 \quad and \quad Cov(x, y) = \rho_{xy}$$

From here we can define a new isolation prediction, that is linearly uncorrelated from the pid prediction. $y' = y - \rho \times x$. Which given $Cov(x, y - \rho x) = Cov(x, y) - \rho Cov(x) = 0$ holds true. In the end we get a new isolation prediction of the following formula:

$$iso' = iso - \rho_{iso,pid} \frac{\sigma^2_{iso}}{\sigma^2_{pid}} \times pid \tag{8}$$

where iso and pid refers to the predictions of the respective models. This whole framework works optimally on gaussian distributions, since the variances and uncertainties assume gaussian distributions. A logit transformation is applied to the raw predictions of our algorithms, this spreads them out beyond the usual range of $[0, 1]$. This transformation is also applied by ATLAS and provides more gaussianly distributed predictions.

$$TML = -log(1/ML - 1)/k \tag{9}$$

Where ML refers to the output of a discriminator in range $[0, 1]$ and TML is the transformed scores. k is an artibitrarily chosen factor, ATLAS uses $k = 15$, however, for this work $k = 10$ has been used.

For a well trained algorithm, this works nicely and provides two well separated gaussian distributions of signal and background. However, while isolation and identification in this framework are two pretty similar tasks (done on a separate set of variables) the outputs vary quite significantly in scale, the correlation and variances will therefore potentially be very different between the signal and background distributions. This is dealt with by calculating the correlations and variances of the signal and background distributions separately and decorrelate using the average, such that:

$$\sigma^2_{pid} = \frac{1}{2}(\sigma^2_{pid,sig} + \sigma^2_{pid,bkg}) \quad and \quad \sigma^2_{iso} = \frac{1}{2}(\sigma^2_{iso,sig} + \sigma^2_{iso,bkg})$$

Applying the DeCorr Framework to Electrons

Given the very different outputs of an electron *T&P* and a photon *TT&P*, the framework takes a slightly different shape, with the electron being by far the simpler example. This is due to the fact that the *T&P* selects a massive amount of signal electrons while selecting a

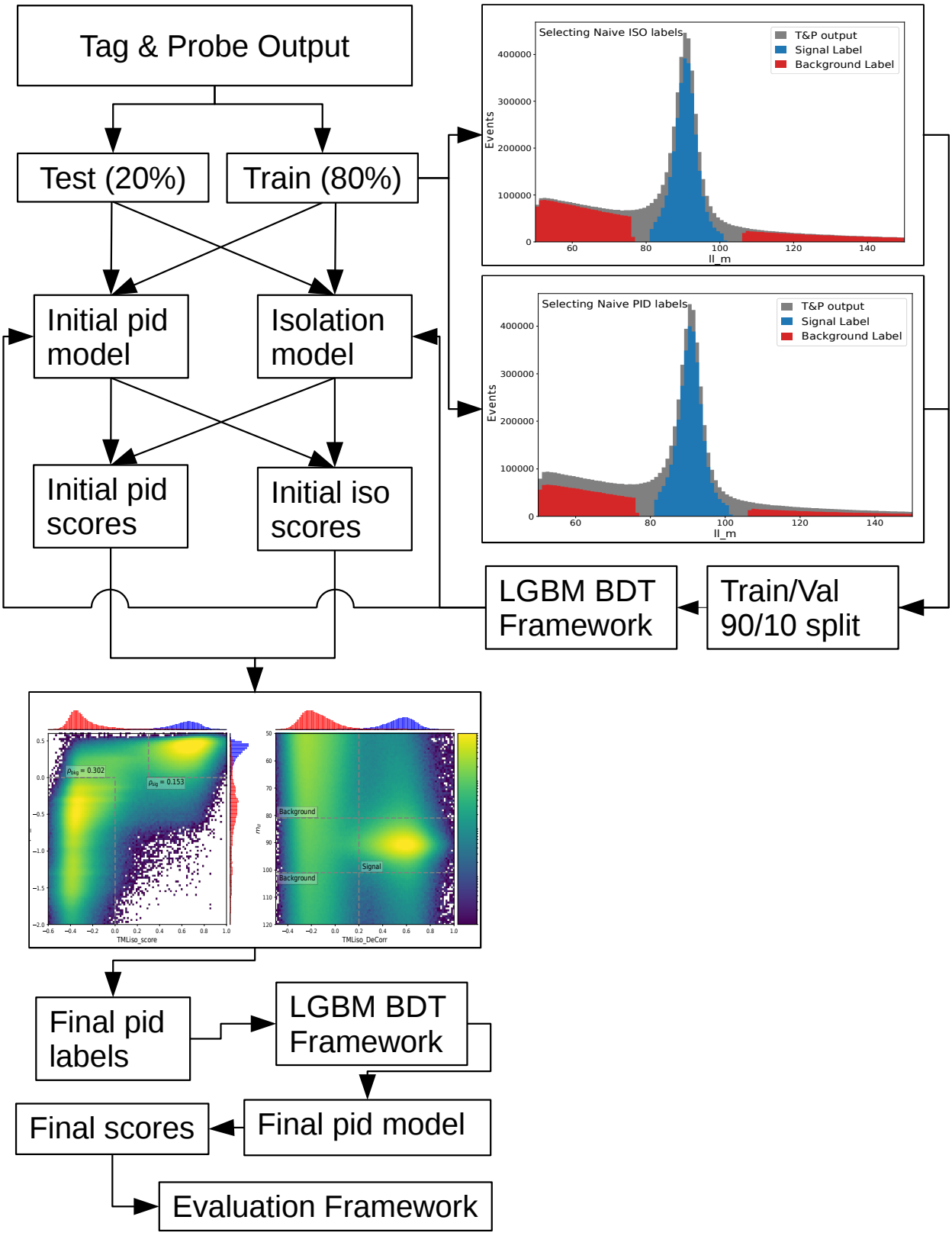


Figure 46: The work flow of the electron data training, decorrelation, training and evaluation. The diagram describes the handling of data as well give context to the plots shown throughout the section.

nicely distributed background that allows for a strong initial label selection. The isolation and identification model predictions acquired from these initial labels are used to create a decorrelated isolation prediction, which is then immediately applied to select new labels used to re-train an identification model. For a full overview of the process see figure 46. The first step is to set 20% of the *T&P* output aside for evaluation. The remaining 80% of the output is used for training. As seen in the top right of figure 46 the ll_m distribution of the training set is shown in grey, while the initial labels for identification and isolation are shown in blue and red, for signal and background respectively.

Task	Signal Selection	Background Selection
Identification	$\lvert m_{ll} - m_Z \rvert < 10GeV, \quad \frac{etcone40}{E_T} < 0.06$	$\lvert m_{ll} - m_Z \rvert > 15GeV, \quad \frac{etcone40}{E_T} > 0.15$
Isolation	$\lvert m_{ll} - m_Z \rvert < 10GeV, \quad Loose$	$\lvert m_{ll} - m_Z \rvert > 15GeV, \quad !Loose$

Table 15: The signal and background selections applied to the probes in order to select initial identification and isolation labels for electron training in data.

The selections used to select these labels are described in table 15. On these labels an isolation model using the isolation variables described in chapter and an identification model using the extended variables described in chapter **??** are trained. The predictions of these models are then used to create a decorrelated isolation prediction. Note that the evaluation set along with predictions of these two models are saved for final evaluation and are not used in any way before then. Figure 47 illustrates the two major steps taken to

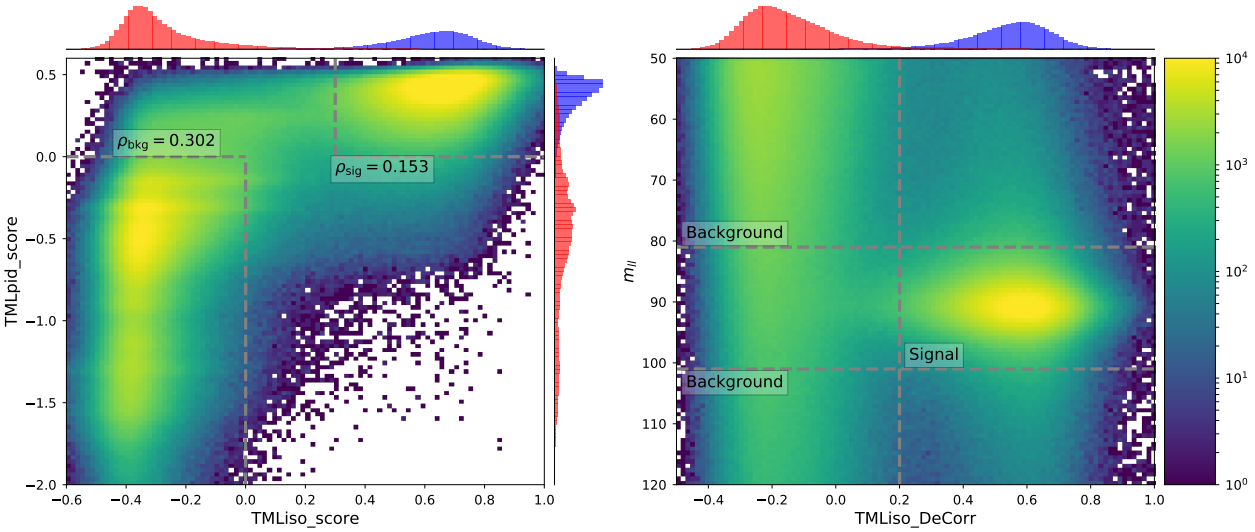


Figure 47: The two plots visualizing the correlations (left) and label selection (right) done on in order to select clean labels in data for electrons. The signal (blue) and background (red) on the projections are aquired by requiring the loose working point to select the object. The correlations listed on the left plot are calculated inside the boxes surrounded by the dashed grey lines. Similarly the label selections are indicated by dashed grey lines, with the

decorrelate the isolation variable and use it, along with the invariant mass m_{ll} to select cleaner labels. The left plot in the figure shows the logit transformed (eq. 9) isolation and identification scores trained on the initial labels plotted against one another in a 2d histogram. The 1d projections of the x- and y-axis has separate signal (blue) and background (red) distributions indicated by use of the loose working point. They clearly show the very nice seperation in both variables, with the isolation score performing the best. This provides a great separation and probing the correlation of the individual distributions

is easily done. The correlations are calculated inside the boxes indicated by the dashed grey lines and stated on the figure, $rho_{sig} = 0.153$ and $rho_{bkg} = 0.302$. These are two very compatible correlations and the average decorrelation will work nicely, as it will slightly overcorrect the signal objects, while slightly undercorrecting background objects. One should expect a final correlation after decorrelating of $\sim \pm 0.075$ for background and signal respectively. These correlations are acceptable and labels are selected using the decorrelated isolation variable as described in the second plot of figure 47. The cuts applied to select signal and background and indicated by the grey dashed lines, where only boxes containing either a *"Background"* or *"Signal"* label will be kept. Deeming from the plot alone, these labels will be extremely clean. This selected signal and background, in addition to the background selected by the background selection described in chapter is passed on as final labels to the training framework. Once again several models with increasing complexity were

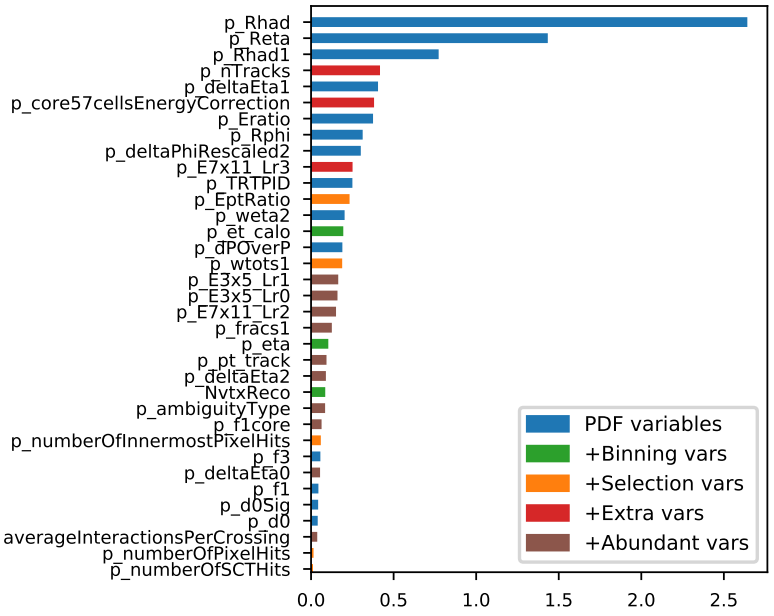


Figure 48: The SHAP value ranking of a model trained on the extra dataset. Note the different variables used for electrons.

trained and the resulting SHAP value ranking and roc curves are shown in figures 48 and 49. Much akin the MC photon training described earlier, a similar approach to the variable search was applied. Initially, a large set of variables were used and quickly a select few additional variables were selected. For electrons, this finalized list of variables is the abundant variable list shown in figure 48. As the name suggests, this list was never intended to be included for an eventual launch. However the three variables included in the extra variable list is a discovery unique to this data training and show promising improvements. This is due to the fact that these variables never ranked anywhere as highly in the MC training done by other students. This provides a wonderful insight into the usefulness of training in data, as the possibility to discover previously underesti-

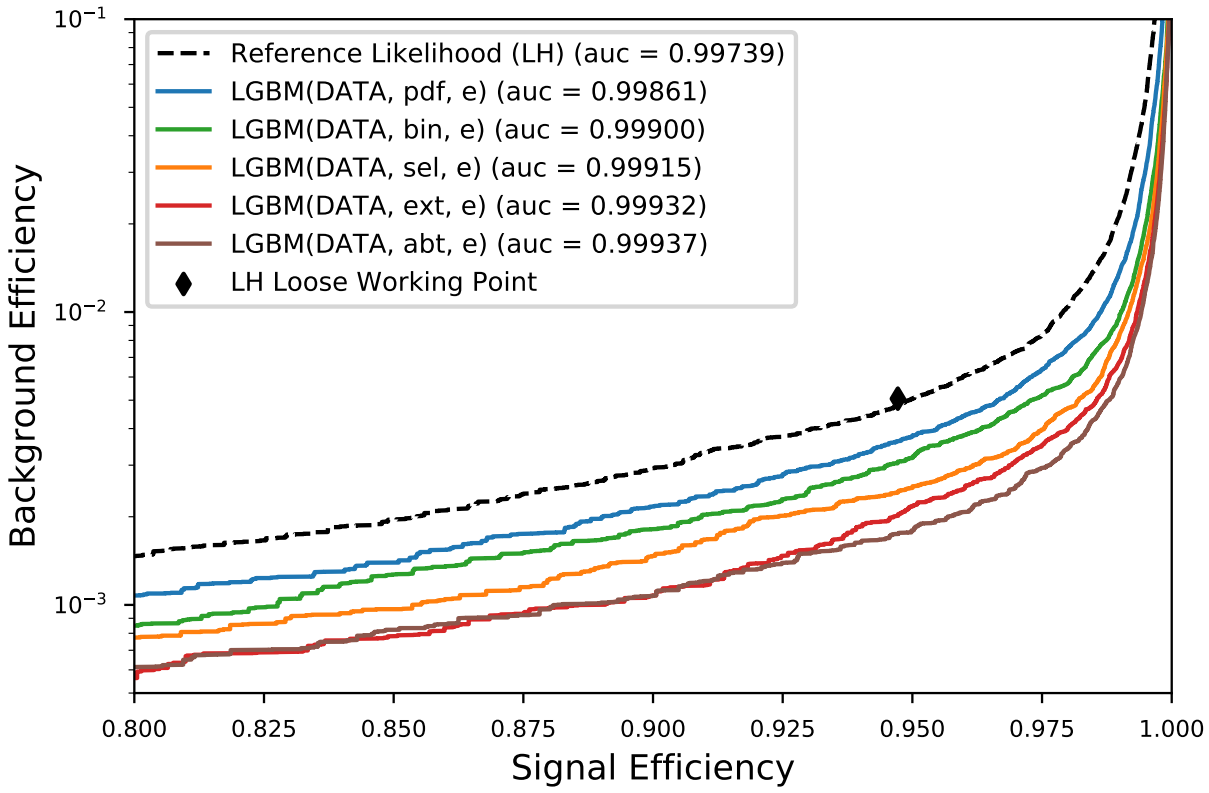


Figure 49: The roc curves of the models trained for each electron variableset on the electron dataset.

mated variables, one would otherwise miss, is an enticing prospect. For this work, especially, considering electrons were done as a case study for the DeCorr framework. The variables used for the selection variable set are the variables onto which the ATLAS electron ID applies cuts, as described in chapter . For a much more complete look at electron identification one should read the work done by another student in the group[**ehrke_machine_2019**]. Nevertheless the electron data training has been examined to an extent that would be sufficient to deem the training successful and the resulting plots can be found in figure 50. These plots correspond to the photon MC evaluation. First column is the background efficiency when matching the signal efficiency of the loose working point, the middle column is the signal efficiency of the loose working point, and the last column is the background efficiency of the models with a fixed signal efficiency of 96%. The rows representing the binning in $\langle\mu\rangle$, $|\eta|$, and E_T respectively. The overall take is that the performance increase is promising, with no real surprises. The increased performance (especially at low energy) gained from the extra variables warrants a further analysis. However since the main aim of this work is to provide photon identification models, the focus will shift back to photons, specifically to applying the DeCorr framework to them.

Applying the DeCorr Framework to Photons

The photon $TT\&P$ selection described in chapter provides the best available method for selecting unbiased photons in data. However, as figure 52 showing the output clearly reveals, the output contains a lot of background. The plot is a 2d histogram of the invariant mass of the tagpair and the probe $m_{ll\gamma}$ and the invariant mass of the tagpair m_{ll}. Probes originating from a $Z \to ll\gamma$ will naturally collect in a resonance around the Z-pole, while the tagpair will have an invariant mass in the range $50 - 80GeV$. Another significant portion of the output contains tags originating from $Z \to ll$ events, where most likely a hadron[22] has been selected as the probe. These events will naturally collect at the Z-pole in the tagpair invariant mass spectrum. This is intended behaviour as the rarity of the signal event and the similarity between the selections forces one to select $Z \to ll$. This provides us with plenty of background and with access to the two invariant masses along with isolation and particle identification we can acquire initial labels. The selections used to acquire initial isolation and identification labels in the $TT\&P$ output are described in table 16. Having acquired initial labels, the next step is to crank the

[22] It could be anything, that makes it into the photon container

Task	Signal Selection	Background Selection
Identification	$\lvert m_{ll\gamma} - m_Z\rvert < 5GeV$, $m_{ll} < 82GeV$, $\frac{ptvarcone30}{E_T} < 0.06$	$\lvert m_{ll\gamma} - m_Z\rvert > 5GeV$, $\frac{ptvarcone30}{E_T} > 0.22$
Isolation	$\lvert m_{ll\gamma} - m_Z\rvert < 5GeV$, $m_{ll} < 82GeV$, $Loose$	$\lvert m_{ll} - m_Z\rvert > 5GeV$, $!Loose$

Table 16: The signal and background selections applied to the probes in order to select initial identification and isolation labels for photon training in data.

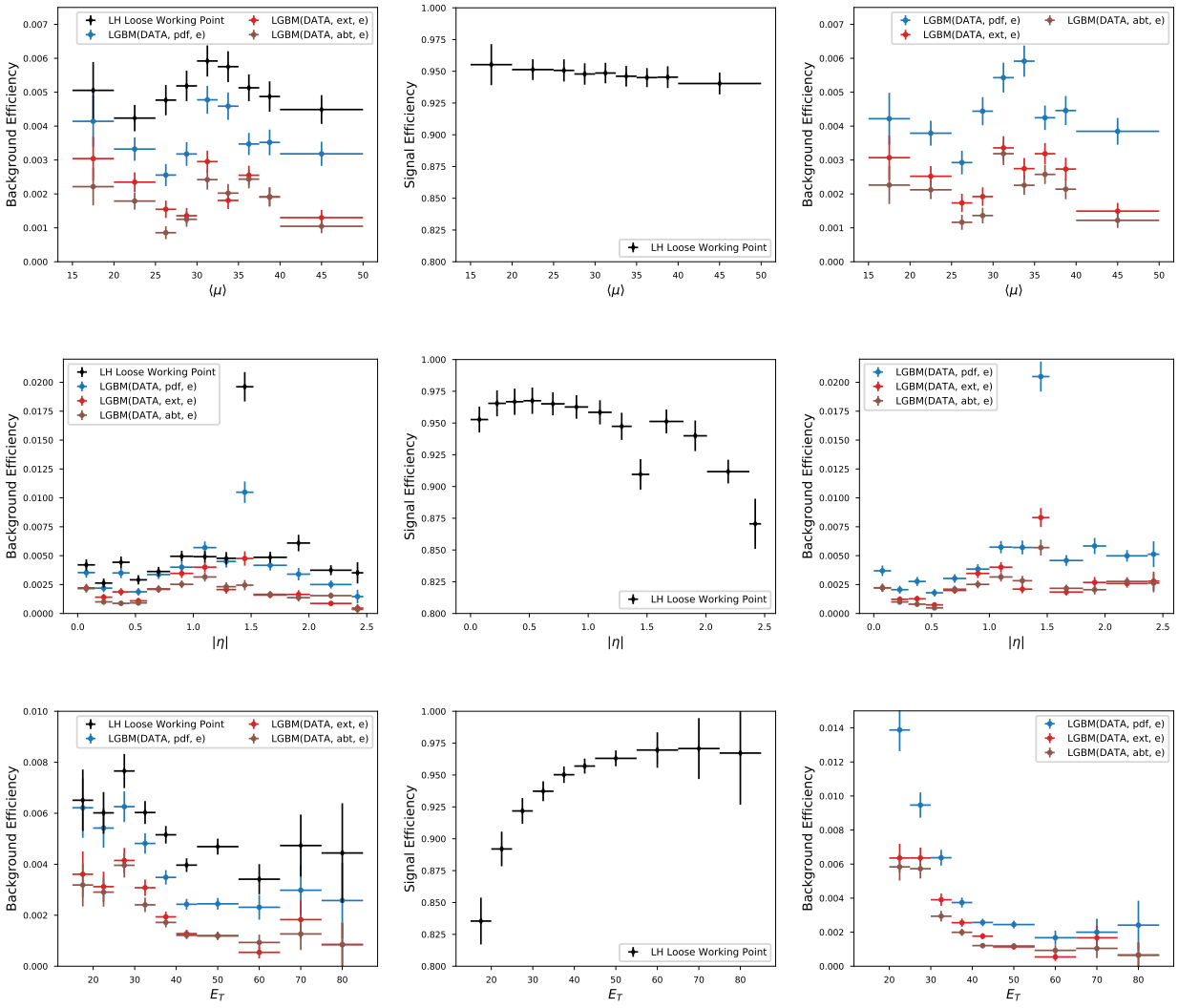


Figure 50: The electron performance plots, all as a function of $\langle\mu\rangle$, $|\eta|$, and E_T. The first column shows the background efficiency of the loose working point and the models with the signal efficiency matched to the one of the working point. The middle column is the signal efficiency of the loose working point. The final column is the background efficiency of the models with a fixed signal efficiency of 96%.

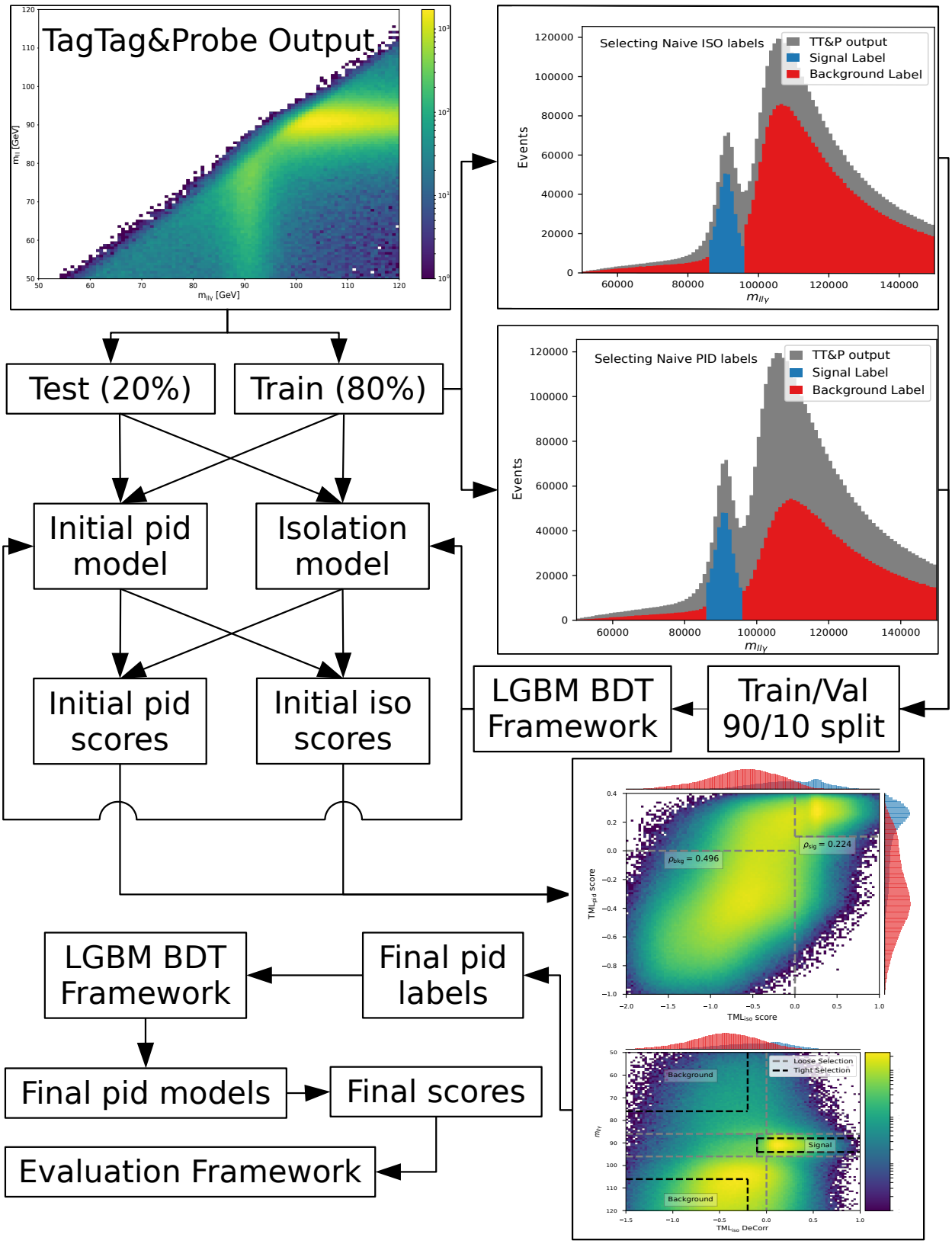


Figure 51: The work flow of the photon data training, decorrelation, training and evaluation. The diagram describes the handling of data as well give context to the plots shown throughout the section.

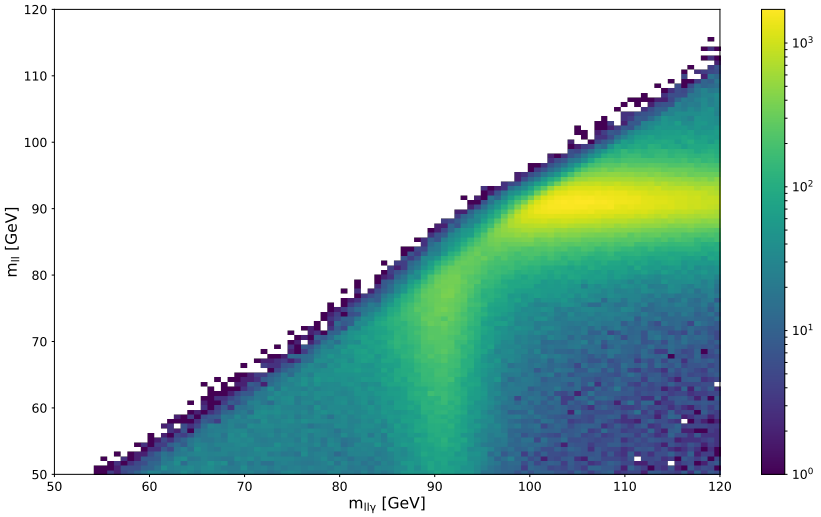


Figure 52: The output of the photon Tag, Tag & Probe, designed to select photons from $Z \rightarrow ll\gamma$ events. The axis are the invariant mass of the tag, tag and probe and tag, tag particles. This shows that the datafiles contains two main sources of probes, actual $Z \rightarrow ll\gamma$ with a high certainty of a signal photon probe and $Z \rightarrow ll$ events, where the probe is a random particle that matched the low requirements of the probe.

wheel of the DeCorr framework, which has to take a slightly different shape due to the more complicated nature of the channel.

A flow diagram of the photon decorrelated workflow is shown in figure 51 on page 86. The overall workflow is very similar to the one found in the electron case, however, additional steps must be added to deal with the increased complexity, these will be described in detail later. Once again, the data is split into train and test data (80%/20%). The test set is set aside and only touched when predicting with models for final evaluation. The initial labels described in table 16 are selected from the train set. In the top right of figure 51, the signal and background labels are visualized along with the $m_{ll\gamma}$ distribution. The full distribution is shown in grey and is the same between the two plots. Given the low statistics of the signal peak and the clear indication that background is still present, these initial labels are seemingly rather confused. The two datasets corresponding to the initial background and signal labels are then split into train and validation sets (90%/10%) and reweighted in $\langle\mu\rangle$, η, and E_T. This is passed to the LGBM framework described in chapter . This returns the initial isolation and identification models, which allows for the application of the DeCorr framework. The logit transformed predictions of the initial models are plotted against one another in figure 53, where the 1d projections of both axes are seperated in red and blue using the loose working point. Comparing to the electron case from figure 47 it becomes apparent that the separations of both models are much worse. This is here the already mentioned increased complexity of the photon case becomes very apparent. For the decorrelation to work, it is vital that we measure the correlation of the signal and background distributions separately. A first attempt at drawing boxes around the distributions is shown as the dashed grey line with the correlations written beside the box. Where $\rho_{sig} = 0.224$ and $\rho_{bkg} = 0.496$. In order to better determine the correlation of the distributions, a signal and background selection is

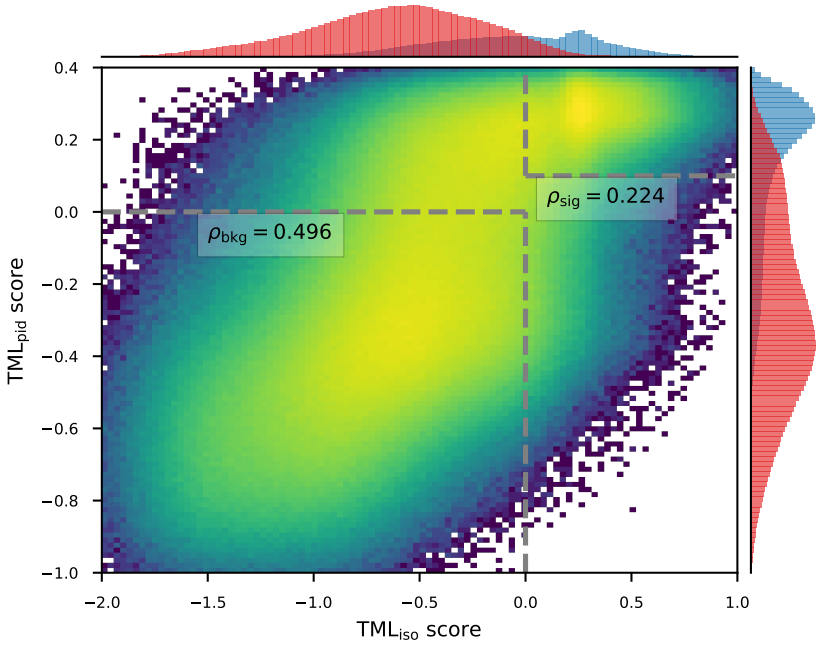


Figure 53: The logit transformed isolatio and identification scores plotted against one another in a 2d histgram. The 1d projections are separated in signal (blue) and background (red) distributions by use of the loose working point. Due to poor separation drawing around signal and background becomes hard. The correlations listed are calculated within the grey dashes boxes drawn, which represent an attempt to probe the correlations of the background and signal distributions.

applied, with the simple goal of isolating the signal and background objects. These selections are the following:

$$Signal : m_{ll} < 83GeV, \quad 86GeV < m_{ll\gamma} < 96GeV \tag{10}$$

$$Background : m_{ll\gamma} < 81GeV \quad | \quad m_{ll\gamma} < 101GeV \tag{11}$$

Applying these cuts and replotting the isolation and identification

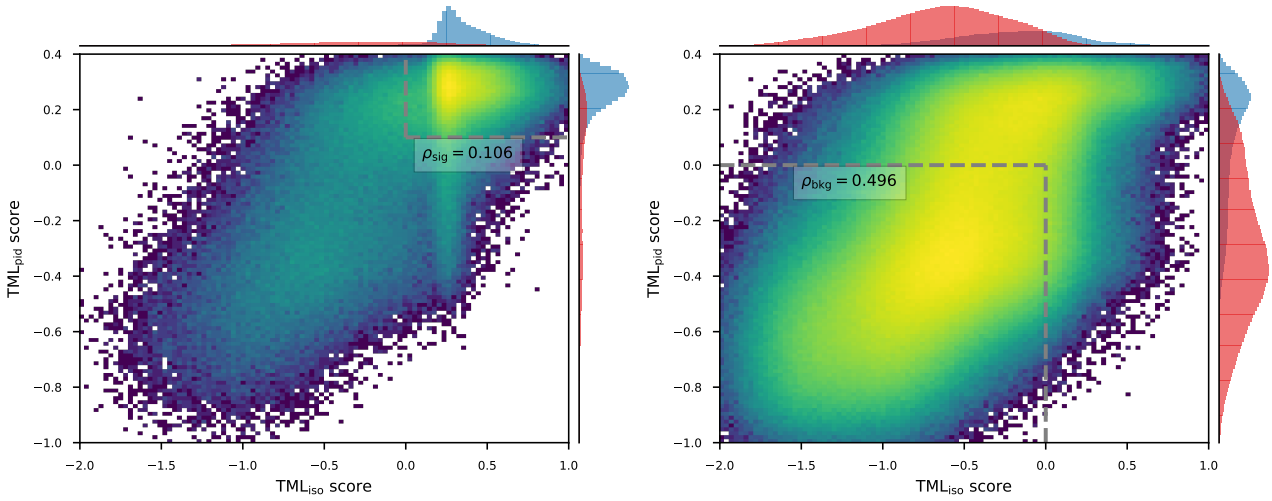


Figure 54: The two plots show a 2d histogram of the logit transformed isolation and identification score distributions, with the distributions projected onto the 1d histograms. The loose working point has been applied to represent signal (blue) and background (red) distributions in the 1d histograms. Signal (left) and background (right) has been selected using requirements in m_{ll} and

scores yields figure 54, where the signal and background selections are applied and shown in the left and right plot, respectively. The signal selection isolates the signal very well, and the measurement of correlation changes to $rho_{sig} = 0.106$. It is much harder to determine the degree of isolation achieved in the background case, but given that the isolation measurement did not change, this measurement is

used as ρ_{bkg}. The distribution of the background selection point to there being real photons in the background, especially given how well the loose working point agrees with the pid score.

The size of the correlations dictates that decorrelating using the average method described in earlier would overcorrect signal and undercorrect background to an extent where the remaining correlations are quite high ($\sim \pm 10 - 20\%$). While this is much better than the alternative, it is suboptimal and other methods have been tested. These methods have been named continuous, piecewise and sigmoid decorrelation.

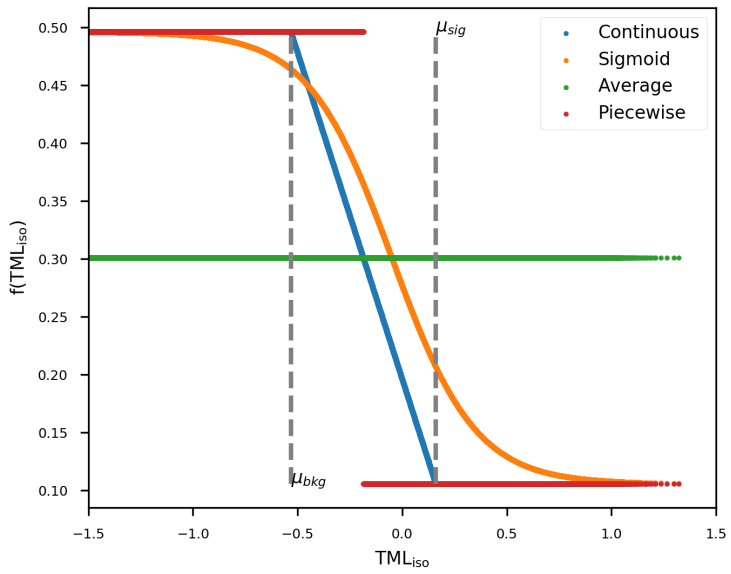


Figure 55: The various methods used for decorrelating, $f(TML_{iso})$ represents the ρ and σ^2 used to decorrelate, which instead of being scalars are now designed as a function of the logit transformed isolation score. This function can follow the evolutions shown in this figure.

Decorrelation Methods All the decorrelation methods, including the already introduced average, has been visualized in figure 55. The task is to find a way to compensate for the large discrepancy in the signal and background correlation.

The piecewise method as visualized infers using ρ_{bkg}, σ^2_{bkg} to decorrelate background objects and ρ_{sig}, σ^2_{sig} to decorrelate signal objects, where signal and background objects are seperated by a cut in isolation. In the best of circumstances, this will perfectly decorrelate, however the poor separation of signal and background in the isolation score has potential to cause discontinuity in the region around the isolation cut, especially due to the large change in correlation.

The continuous method attempt to deal with the possible discontinuity by replacing the hard cut in isolation with a linear transition between ρ_{bkg} and ρ_{sig}. This entails defining the following functions:

$$f(iso) = \alpha iso \times \beta \quad \alpha = \frac{x_{sig} - x_{bkg}}{\mu_{sig} - \mu_{bkg}} \quad \beta = x_{sig} - \mu_{sig} \times \alpha \tag{12}$$

where x refers to the correlation ρ and the variances σ^2_{pid} and σ^2_{iso}.

μ_{sig}, μ_{bkg} refers to the mean of the isolation score of the signal and background distributions, respectively. The functions then replace the $\rho_{pid,iso}$, σ^2_{pid}, and σ^2_{iso} in equation 8.

The sigmoid method has a similar motivation to the continuous and is an attempt to fully avoid discontinuity by approaching the change in correlation in the smoothest way possible. The equation is the following:

$$f(iso) = x_{sig} + \frac{x_{bkg} + x_{sig}}{1 + e^{(iso+\mu)/\sigma}} \quad \mu = 0.05 \quad \sigma = 0.20 \tag{13}$$

where x again refers to the correlation ρ and the variances σ^2_{pid} and σ^2_{iso}. μ and σ was originally defined as $\mu = \frac{1}{2}(\mu_{sig} + \mu_{bkg})$ and $\sigma = \frac{\mu_{sig} - \mu}{2}$. However, the values from equation 13 were found to work the best. The function is purposefully shifted to the right of the intersection between continuous and average to compensate for the larger spread in isolation of the background distribution. It is simply more rewarding to focus on correctly decorrelating the background.

The decorrelation achieved by all methods is summarized in figure 56, where the original distribution is shown in the first column, followed by the *average*, *piecewise*, *continuous* and lastly the *sigmoid*. The x-axis is the isolation score decorrelated by the respective method. The rows show the total distribution, the signal distribution, and the background distribution in order. The gray lines indicate the cuts within which the correlations are calculated. Given the positive correlation a decorrelation can be visualized as a shift in the positive direction in the TML_{iso} score. This shift is much larger for objects with a lower TML_{pid} score. This clearly expresses the overcorrection applied to signal by the *average* method, which produces a clear negative correlation and as expected a insufficient decorrelation has been applied to the background. The resulting correlations for *average* are $\rho_{sig} = -0.125$ and $\rho_{bkg} = 0.209$. Both the *piecewise* and *continuous* methods clearly produce an unfortunate discontinuity, which removes them from contention. However, the *sigmoid* method produce an almost perfect decorrelation with no discontinuity. It has a similar overcompensation of the signal distribution to the *average*, however, it contributes with a much smaller correlation.

The conclusion is that one can achieve an isolation score that has a significantly reduced linear correlation with identification using the *average* method and an isolation score that is almost entirely linearly uncorrelated with identification using the sigmoid method, from which the resulting correlations are found to be $\rho_{bkg} = 0.002$ and $\rho_{sig} = -0.021$.

Unfortunately due to a late addition of the sigmoid method, the labels were selected using the *average* decorrelated isolation score. However, for reference the label selection with the sigmoid decorrelated isolation score can be seen in appendix , which shows that the effective difference in the labels are small and could be corrected for by applying a slightly different cut.

The label selection is done by cutting in the decorrelated isolation

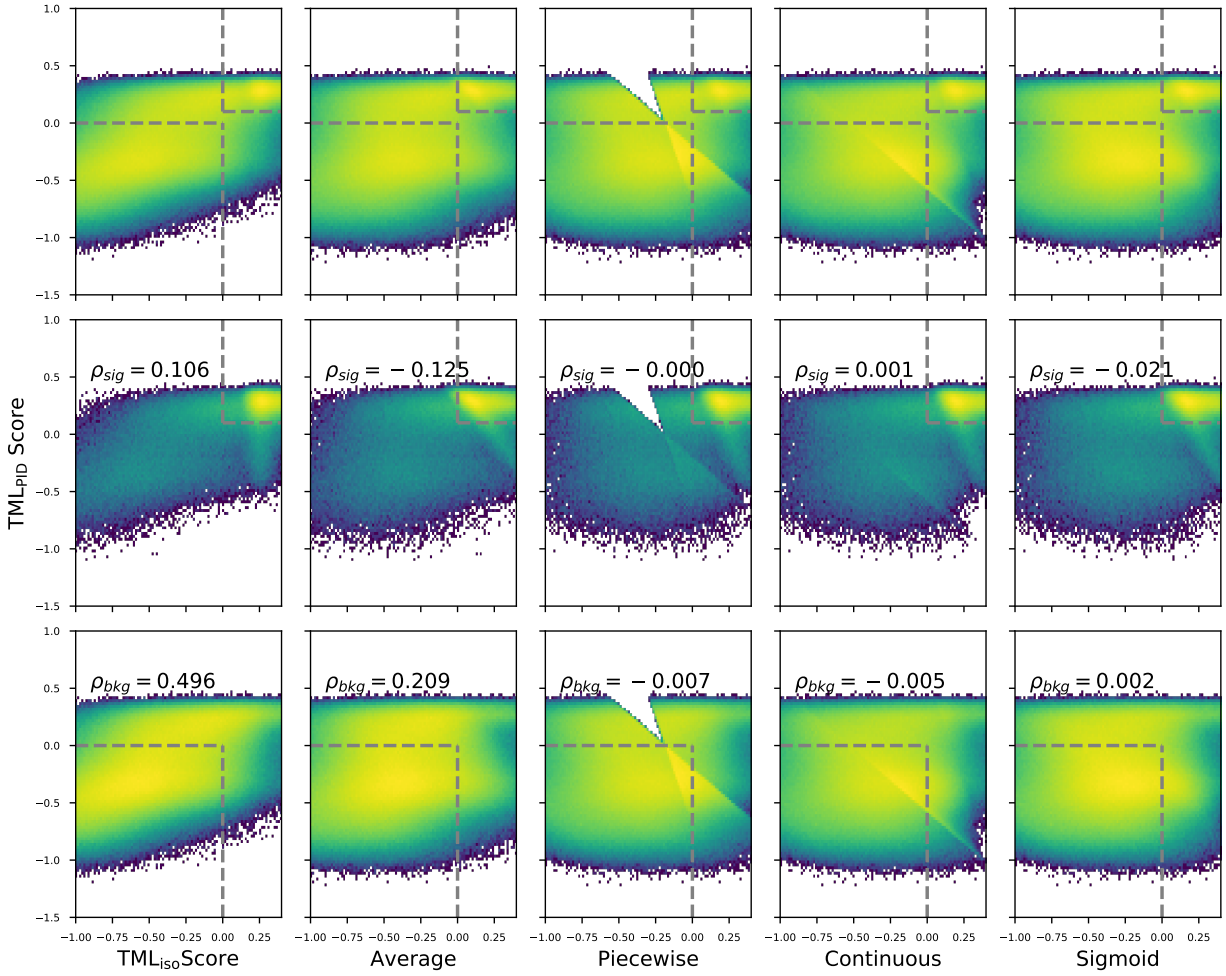


Figure 56: The first column here shows the three correlation figures already shown in this section. Namely to total distributions of logit transformed isolation and identification scores in a 2d histogram, the selected signal distribution, and the selected background distribution. The correlations listed and the ones calculated in the boxes drawn. The following columns show the same plots, with the decorrelated logit transformed isolation score on x-axis. The method used for decorrelation is indicated by the x-axis label. The degree to which decorrelation is achieved is evaluated using this figure.

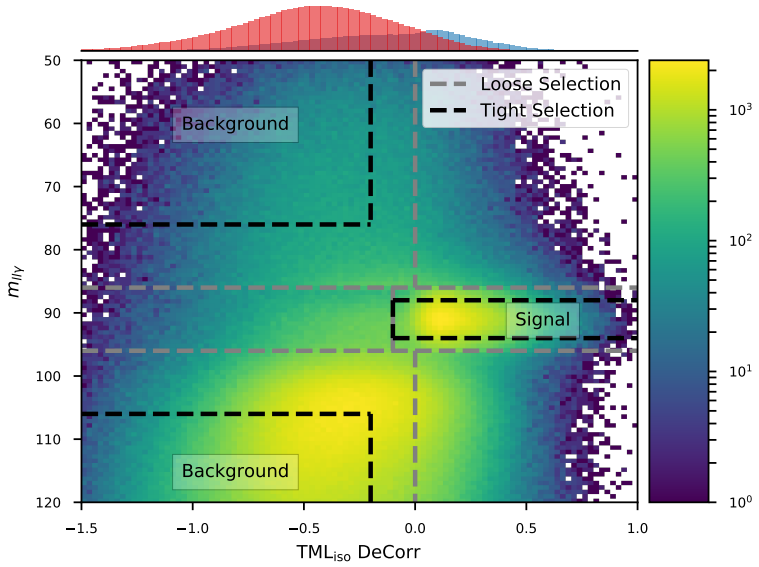


Figure 57: A 2d histogram of the average decorrelated logit transformed isolation scores and the invariant mass $m_{ll\gamma}$, with the x-axis projected into the 1d histogram, where the loose working point has been used to separate the signal (blue) and background (red) distributions. The separation in this 2d space allows one to select signal and background labels. This is done in two ways, a loose and a tight selection. The latter is a subset of the former. Only the boxes containing a *"Signal"* or *"Background"* label are kept.

score referred to as TML_{iso} DeCorr and in the invariant mass $m_{ll\gamma}$. The label selection is summarized in figure 57, where the 1d projection of the decorrelated isolation score shows how the separation of the variable alone is rather poor. Note the colors red and blue once again represent passing the loose working point or not. However, with use of the invariant mass, one is clearly able to separate signal from background to a satisfying degree. Two selections are applied, indicated by the gray and black dashed lines, named loose and tight. The tight selection is a subset of the loose selection and only boxes containing a *"Background"* or *"Signal"* label are kept, the rest is not kept for training. This results in two new datasets available for training, both of which are put through the full LGBM framework.

Additions to Photon Data Training A few things need to be adressed before evaulating the data photon training. Firstly, the naming scheme introduced in equation 7 requires an addition. This is due to the multiple label selections used in data. Models are trained on three datasets, each differentiated by a unique label selection. These are the initial, loose and tight labels. Where the loose and tight labels are selected using the decorrelated isolation score. This will be specified in the following expansion of the model naming scheme:

$$Modelname : "MODELTYPE"("DATATYPE","varlist","labelselection") \tag{14}$$

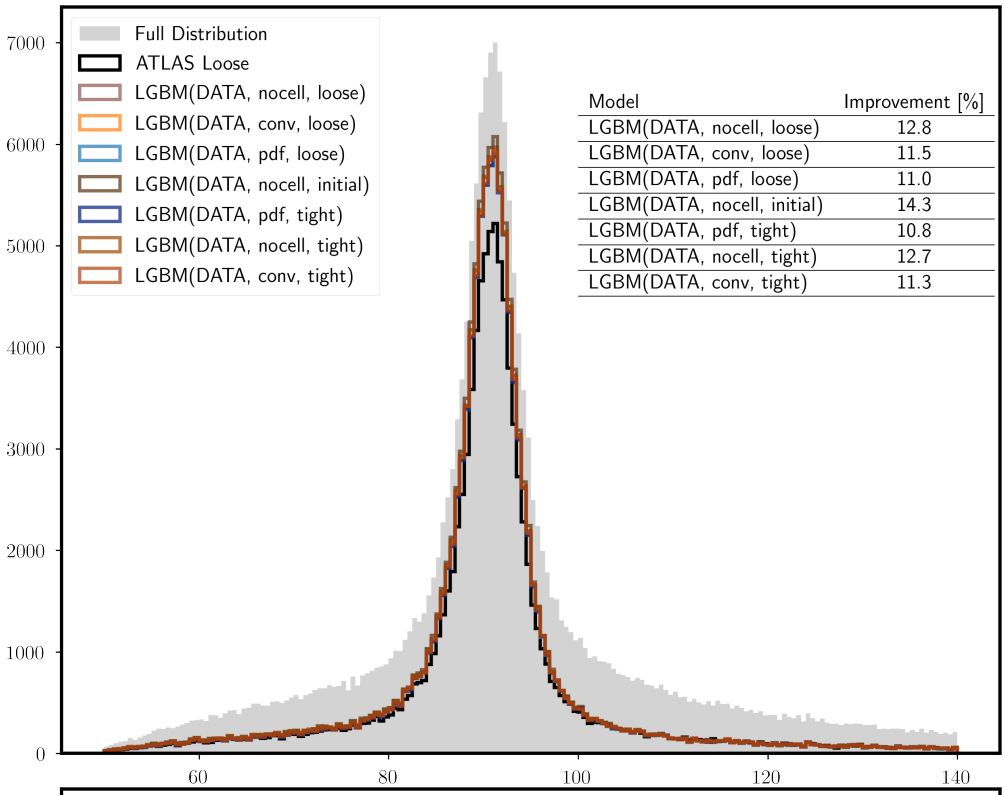
Resulting in a LGBM model trained on the tight label selection, using the conversion variable set will be named $LGBM(DATA, conv, tight)$. The production of photon data was unfortunately delayed over several attempts and due to the increased need for statistics due to the rarity of the process a decision was made to use the DAOD files

up kept and produced by E/γ using their derivation frameworks introduced in chapter . This caused a list of problems, varying in severity from having to switch back to $AntiKt4EMTopoJets$[23] to entire variables being missing. The latter unfortunately affecting the best performing extra variable set. Forcing the definition of a new variable set called nocell, which contains all the same variables as the extra variable set except for the variables $maxEcell\_time$ and $maxEcell\_energy$.

[23] The in-house derivation framework used the recommended pflow jets

Evaluating Photon Data Performance The evaluation methods applied to MC ultimately rely on the assumption that the labels are true. This assumption was well founded for the electron case, however, quite the opposite is true for the photon case. However, an evaluation available in data that makes no assumptions is fitting of the resonance, one can acquire signal and background efficiencies from fitting the signal and background distributions on the resonance before and after applying a threshold on the models. This is ideally done in bins of η and E_T to allow for direct comparison with the working points. However, the statistics of the sample does not allow for a binned approch and the rather tricky shape of the background distribution, seen in top right corner of figure 51, complicates fitting. For comparison, an electron $T\&P$ the sidebands of the invariant mass plot provide for a nice fit of the background distribution. The complete and proper evaluation was simply outside the scope of this project. A simpler method of evaluation was deemed a workable solution. Where the background efficiency is matched in the region $m_{ll\gamma} > 110GeV$. This allows for a comparison between the amount of signal selected by the models and the working points.

The two performance plots are shown in figure 58, where a select group of the models are compared to the working points, with the improvements listed in the table in the plots. The improvements are calculated by simply comparing the number of events selected. This measure is a good probe for the improvement is signal efficiency, due to the equal background efficiency, which is supported nicely by the sidebands being equal. The worry here is that the models might shape the background distribution into a peak, however, this effect, if present, should be minimized by the reweighting. In general, all models listed out perform the working points. One can once again note the fact that increased model complexity leads to better performance. $LGBM(DATA, nocell, initial)$ performs the best against the loose working point and $LGBM(DATA, nocell, loose)$ performs the best on the tight working point, with the other label selections following close behind. Over all, the performance of the models trained on the loose label selection performs at best similar to the initial training, while the models trained on the tight label selection perform worse. This means that the information lost by cutting on isolation (correlated with PID) to select labels was not significant enough to affect training. However, the fact that the unbiased label selection performs at least equal to the initial labels is very positive,



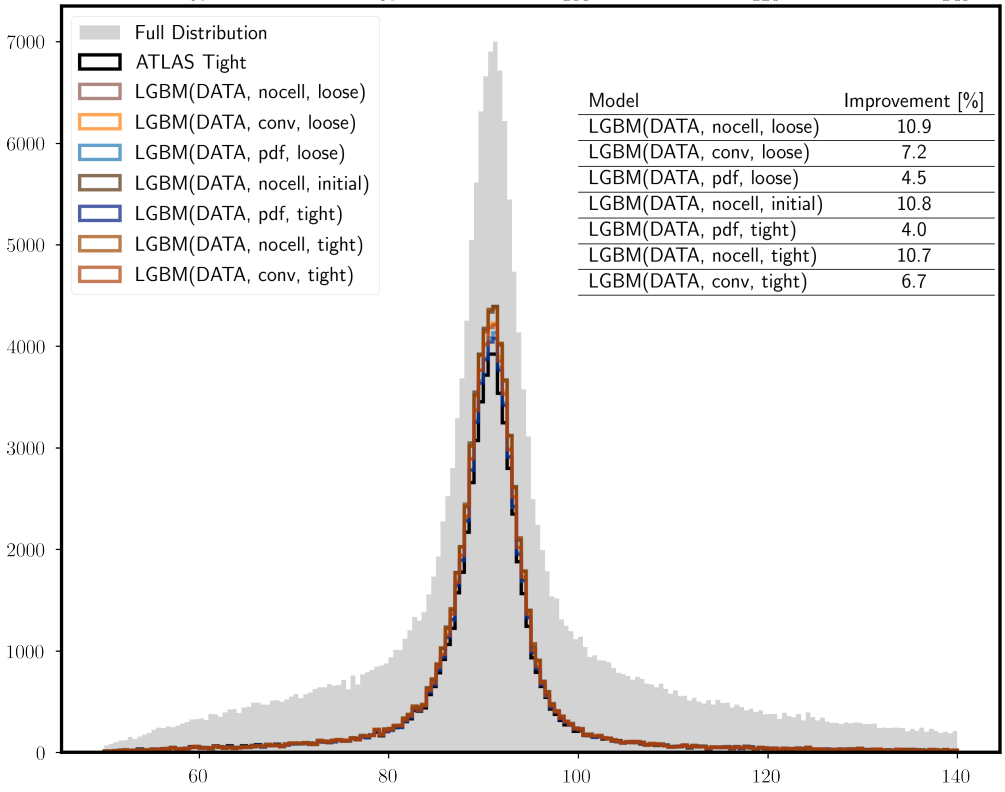
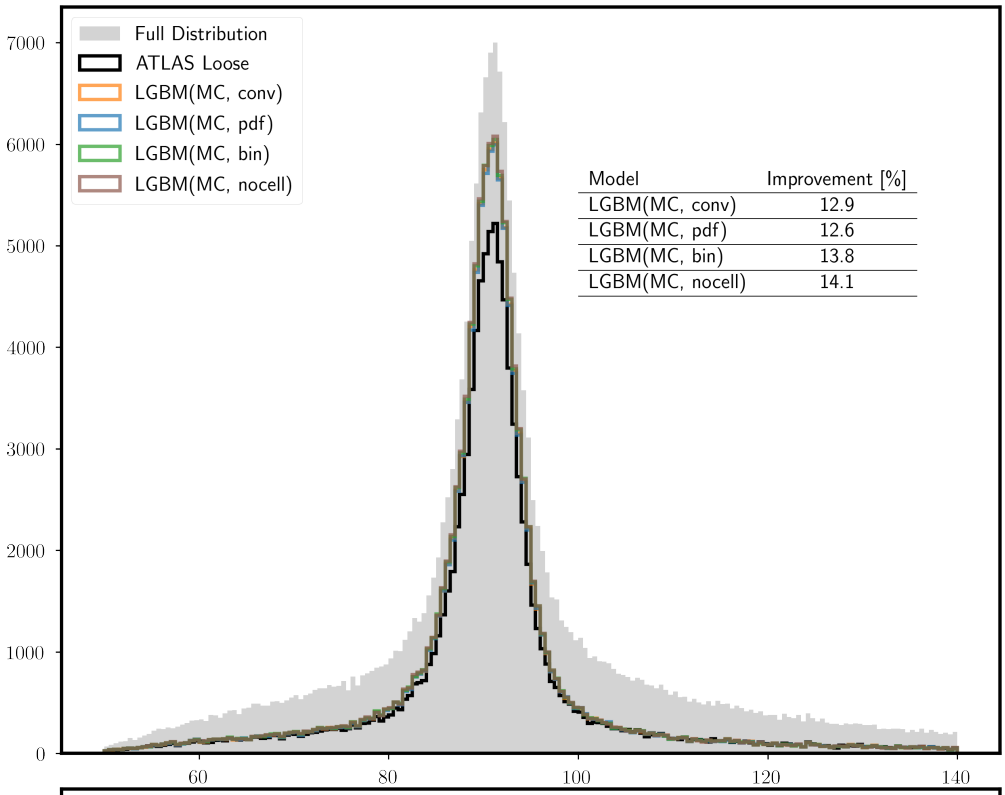


Figure 58: The distribution in $m_{ll\gamma}$ of the evaluation set after requiring $m_{ll} > 82GeV$. The signal efficiency is visualized for the loose (left) and tight (right) working points and the data trained models, where the background efficiency has been matched to the one of the relevant working point in the background region $m_{ll\gamma} > 110GeV$.

given that it should, by design, include fewer potential biases. This was the main goal of the whole method. In addition to this an isolation measurement that is uncorrelated from PID has been gained in the process. In general the training is affected by low statistics and perhaps with higher statistics and a more complete evaluation, one could perhaps find larger differences in the models trained on the different selections. However, the evaluation used seemingly favours the initial and loose label selections.

The MC trained models are naturally also evaluated on the data sample, figure 59, where each of the available models are compared to the loose and tight working points. The improvements are listed in the table in the plots. These results show yet another indication that the MC training, with all its flaws and challenges does exactly what it is trained to do, select photons from objects that behaves similarly to a photon. This improvements are smaller than the data trained models, however, given that it is trained on a data from the whole energy range, including several resonances, it is impressive that is behaves so nicely on this given resonance.

The main point of concern is the evaluation on the overall distribution, is that the improvements could be distributed poorly across phase space. It is therefore hard to say anything conclusive. Other than the improvements shown in data are promising. The fact that the MC trained models, who has been exposed to a much wider range of datapoints than just ones from the Z resonance, also show improvement further strengthens the promise. The stability of the MC trained models across phase space should to some extends reflect the improvements seen in data. However, without applying the binning to data, there is no way to know for sure. This did not manage to fall within the scope of the project and will be left for the next lucky soul, who has the courage to tackle photons in the ATLAS detector.



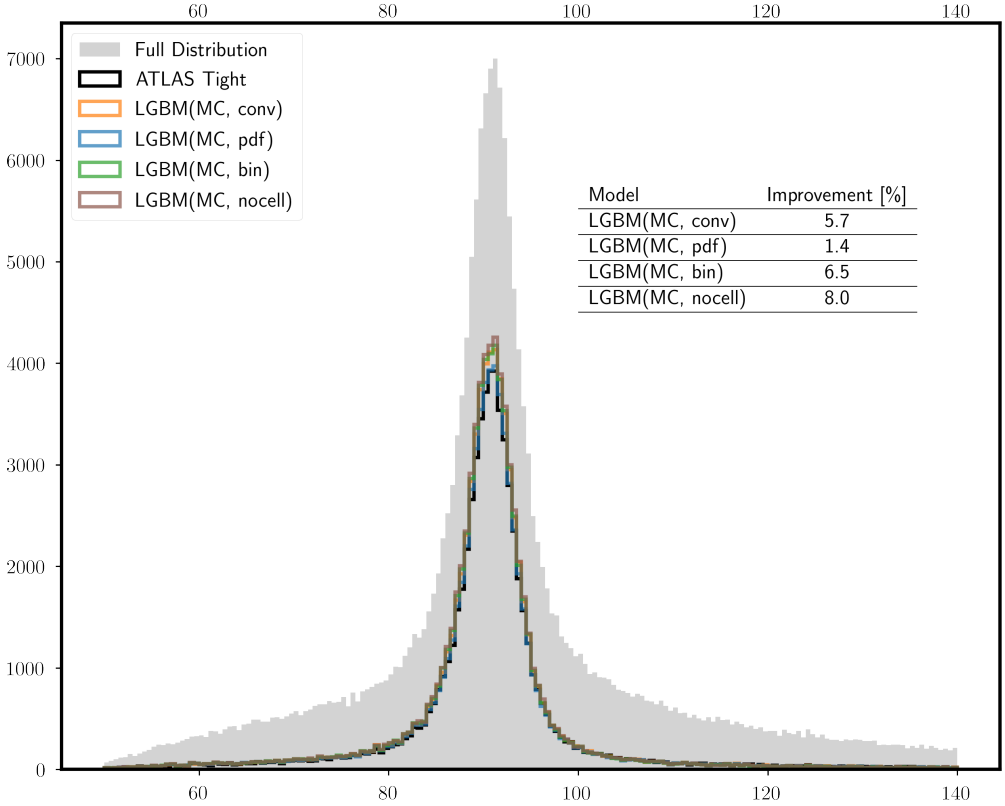


Figure 59: The distribution in $m_{ll\gamma}$ of the evaluation set after requiring $m_{ll} > 82GeV$. The signal efficiency is visualized for the loose (left) and tight (right) working points and the MC trained models, where the background efficiency has been matched to the one of the relevant working point in the background region $m_{ll\gamma} > 110GeV$.

Evaluating on Higgs $\rightarrow \gamma\gamma$

The pinnacle of photon particle identification is selecting Higgs bosons. The channel is known as a high background channel, however, much of this background comes from two actual photons, produced in other SM processes. This means that one's ability to select photons directly propagates through to selecting more Higgs bosons, however, also more background. Thus the net gain from a better identification model is more data (signal + background) of a similar purity, which is an obvious advantage.

The MC trained models are evaluated on the Higgs MC trained dataset described in chapter . As described in chapter , the $H \rightarrow \gamma\gamma$ analysis is very complex and involves further event reconstruction, beyond the scope of this project and the full analysis was not recreated. However as described in chapter , as much of the event selection as possible was applied. It was not possible to perform any background selection why it is therefore impossible to match the background efficiencies of the models and working points using this sample. However, given that we have access to the massive photon MC dataset, that has plenty of statistics across the whole E_T spectrum, the matching can be done using that.

Given the E_T distribution of the photon candidates shown in figure 60, the background efficiency was matched in the $50 - 80GeV$ range and the corresponding cuts from the models were applied to the Higgs sample for comparison with the models. This is without a doubt a favourable comparison for the models, given they are optimized against this background and the already established behaviour of the unconverted channel as a function of E_T, however, it was established that these effects affected the working points in a very similar way to the models and it is therefore not entirely unfair. The resulting Higgs peak seperated in the unconverted and converted channels is shown in figure 61, with the relative improvements shown in the table in the plots. Unfortunately, there was a bug in the calculation of the conversion variables, this affects the results in two clear ways. Firstly, all but the $LGBM(MC, pdf)$ model show huge degradation on the converted channel. However, as variables increase beyond the ones of the conversion variable sets, the models regain some performance. This along with the fact that all the models perform nicely in the unconverted channel, leads to the theory wrongly calculated conversion variables being at fault. This theory was supposed by further analysis. Another smaller effect is

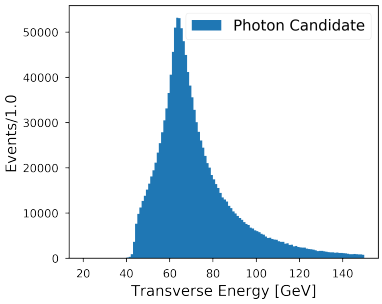


Figure 60: The transverse energy distribution of the photon candidates from the $H \rightarrow \gamma\gamma$ MC dataset.

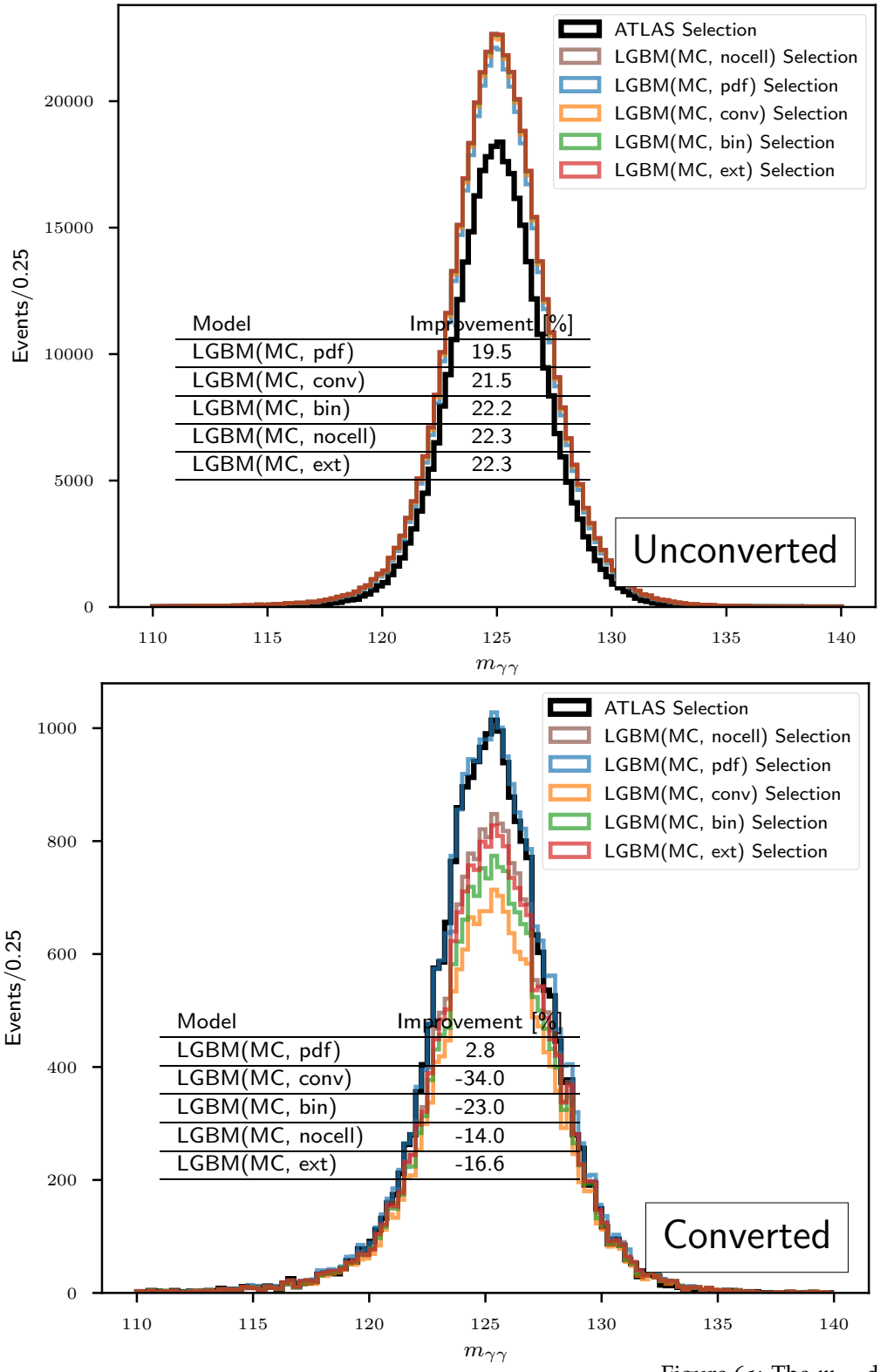


Figure 61: The $m_{\gamma\gamma}$ distribution of the $Higgs \to \gamma\gamma$ MC dataset for the unconverted (left) and converted (right channels).

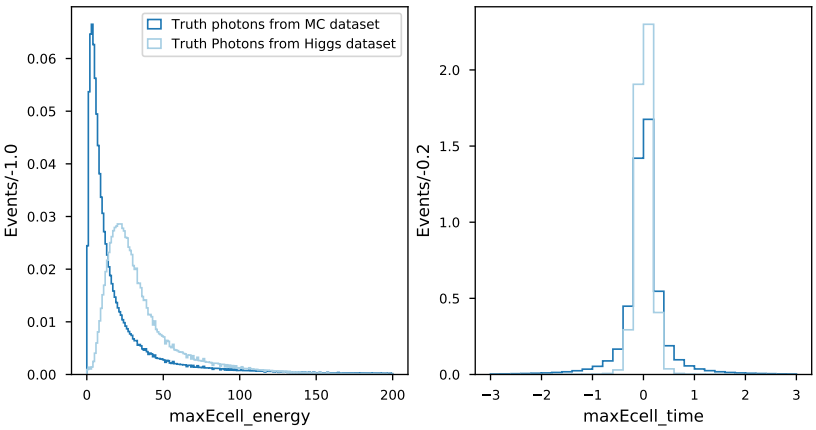


Figure 62: The *maxEcell_energy* and *maxEcell_time* variables, which are the cause of the slight degredation in the $LGBM(MC, ext)$ model performance on $H \to \gamma\gamma$. The plot shown the variable distribution of the MC photon dataset and the Higgs dataset, they histograms have area of 1

the slight degradation of the $LGBM(MC, ext)$ model that performs on par with the $LGBM(MC, nocell)$ model, meaning that no information was gained from the extra variable list. The slight degradation in performance is due to the *maxEcell_energy* and *maxEcell_time* variables being quite different from the MC photon dataset and the higgs dataset, as shown in figure 62.

The overall performance (and degradation) is visualized quite nicely with the distribution of the scores, shown in figures 63 and 64 for the unconverted and converted channel respectively.

While these effects are unfortunate, the performance on the unconverted channel for all models is incredibly promising with improvements upwards of 20%. While this performance is almost certainly overestimated, it is hard to imagine it being overestimated by such a significant margin as for there to be no improvement at all. Even if the $LGBM(MC, pdf)$ model performs equal to that of the ATLAS selection, there is still $2-3\%$ improvement for the more complex models. The improvements shown in MC promise that this is at the very least true. However, it is hard to say anything conclusive about the true improvement on the $H \to \gamma\gamma$ channel until a one to one selection is applied, with background included, while the current results do look promising.

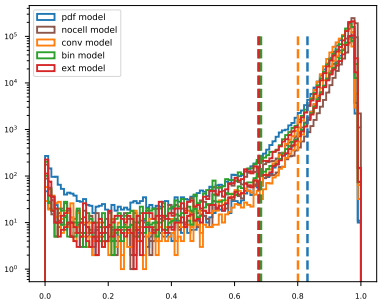


Figure 63: The $H \to \gamma\gamma$ score distribution of the MC models and there corresponding cut (dashed vertical line). Note the small degredation of $LGBM(MC, ext)$

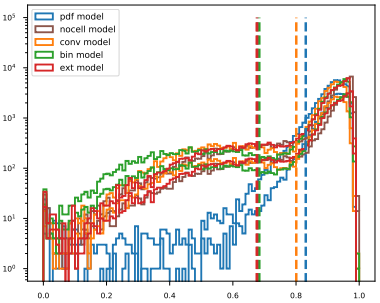


Figure 64: The $H \to \gamma\gamma$ score distribution of the MC models and there corresponding cut (dashed vertical line). Note the large degredation of all models but $LGBM(MC, pdf)$

Summary and Outlook

The aim of this chapter is to summarize the results introduced in the previous chapter, along with the results will be a re-iteration of the caveats and a plan as to how one could test them. The chapter will also include a list of points, which could be addressed in future work on the same, or a similar, problem.

The overreaching conclusions of the work done for this thesis is that the possibility of improving electron and photon identification using machine learning is very real. All the models trained on both particles performed significantly better than the working points in the tests performed and while some of this performance is probably overestimated due to caveats, the overall performance of all trained models look promising. The inclusion of additional variables had proven to consistently increasing performance and is also promising, for the current as well as new methods. The label confusion study clearly showed the robustness of our models and opened up for a long list of interesting questions, many of which were answered. The data-training lead to the discovery of three new variables to potentially include in electron identification.

The improvement in signal efficiency at similar background efficiencies, when evaluating on $Z \rightarrow ll\gamma$ in data is around 14% for both MC and data trained models when comparing to the Loose working point and around 10% for the data trained and 8% for the MC trained models when comparing to the Tight working points. Given the consistency of the improvement shown by the MC trained models when evaluated in both MC and data, one would expect the improvements seen in data to distribute itself similarly to those seen in MC. This could be verified by performing a similar binned comparison. However, this bring up the largest problem with the comparisons done in data. The improvements shown are overall and their distribution across phase space has not been dissected. This must be done before claiming any conclusive improvements and the method for doing so involves acquiring properly measured signal and background efficiencies by fitting.

The evaluation on the MC Higgs peak shows a sizable improvement in signal efficiency, upwards of 22%, in the unconverted channel, while bugs unfortunately did not allow for proper evaluation in the converted channel. This evaluation is also an overall comparison. The matching of the background efficiency is unfortunately done on a background composition that is highly unlike the one used

for optimization of the Tight working point. However, the fact that the negative effects of the background composition clearly affected the working points and the models in the same way is comforting and allows for some confidence in the performance increase being promising. Ultimately, to proper evaluated on the $H \to \gamma\gamma$ channel, one has to do the full selection applied by ATLAS. This allows for a background sample to be selected, in which the background efficiency could be matched.

The isolation model trained on electrons provided a great separation and made for easy use of the decorrelation framework. The same was not entirely true for photons and while the isolation model provided rather poor separation it was sufficient for the decorrelation framework to function. The decorrelation itself revealed two functioning decorrelation methods, one is the simple average that working exactly how expected and the other is the sigmoid method, which provided incredible levels of decorrelation, without producing discontinuities.

Looking onto the future work, beyond what has already been introduced as further tests to the improvements. The main issue, that has not been properly addressed by this work, is the gain from having selections that can be applied to both MC and data. This is true for the selections used to create the electron and photon data datasets. Having corresponding MC sets allows for answers to a lot of the questions that arise from data training. The main one addressed in this thesis is the shaping of the background, this is problem that is assumed to be dealt with by the reweighting, but it is hard to test without MC truth matching. The first step in the future work should be to obtain a MC sample using the photon data selection.

The $Z \to ll\gamma$ was run on all of data17, however more data could be included. This will increase model performance and allow for binned evaluation.

All the models, electron identification and isolation, photon identification and isolation should be implemented in Athena, the software framework of ATLAS. The models are implemented and working in Eventloop, which is a good step on the way. The implementation is a important in order to fully test the models.

Having implemented the models, a great testing ground for the photon identification and isolation models are running the $H \to \gamma\gamma$ selection using the models instead of the ATLAS isolation and identification requirements. In the same vain, one could test both the electron and photon models in the $H \to Z\gamma$ channel[**atlas_collaboration_searches_2017**]. Testing channels like these are the ultimate goal of a new implementation, since this is exactly where the matter the most.

Appendix

Triggers for Electron Selection

Background Prescaled Single Electron Triggers	
HLT_e5_etcut	HLT_e40_etcut_L1EM15
HLT_e10_etcut_L1EM7	HLT_e40_etcut_L1EM15
HLT_e15_etcut_L1EM7	HLT_e60_etcut
HLT_e20_etcut_L1EM12	HLT_e80_etcut
HLT_e25_etcut_L1EM15	HLT_e100_etcut
HLT_e30_etcut_L1EM15	HLT_e120_etcut

Table 17: Background electron triggers, they are all prescaled.

Lowest unprescaled Triggers
HLT_e26_lhtight_nodo_ivarloose
HLT_mu26_ivarmedium
HLT_g300_etcut

Table 18: The lowest unprescaled trigger required of the electrons, muons, and photons

Lowest unprescaled Electron Triggers
HLT_e26_lhtight_nodo_ivarloose
HLT_e60_lhmedium_nodo
HLT_e140_lhloose_nodo

Table 19: The lowest unprescaled electron triggers

Various triggers are required throuhout the electron event selection. One of the triggers found in table 17 is required of the Background Electron pre-selection. One of the triggers found in table 18 has to passed for the $Z \rightarrow ee$ *T&P*. One of the triggers found in table 19 is required of the Z TagElectron pre-selection.

Learning Rate Finder Output

The BDT learning rate finder run on 10% of the MC photon dataset, using the extra variable set.

Sigmoid Label Selection

The 2d mass versus decorrelated isolation score, when decorrelating with the sigmoid method is shown in figure 66. The better decorrelation leads to the the background distribution overlapping slightly more with the signal distribution in the isolation score. Therefore the cut applied to background is a little harsher and sigmoid will

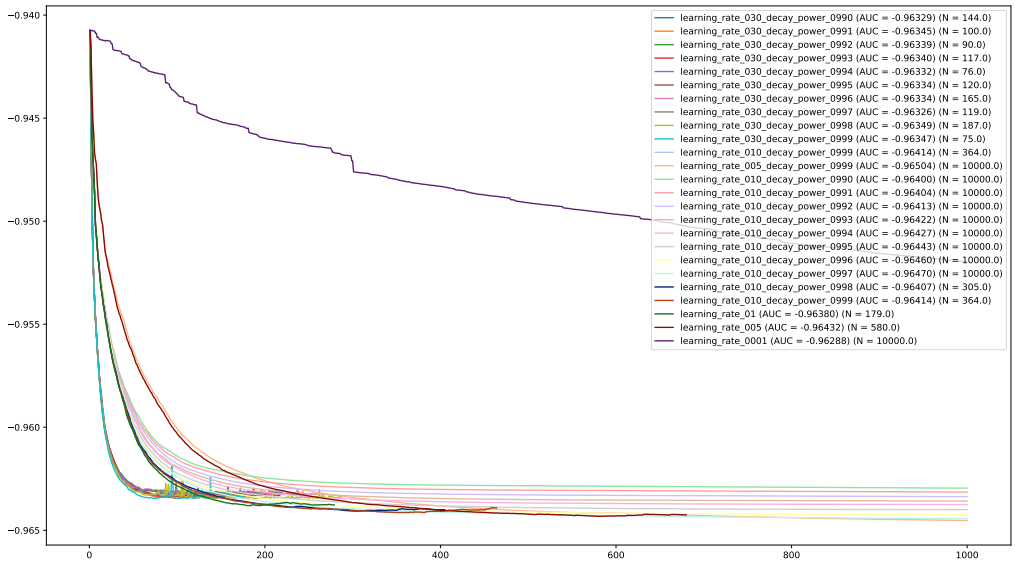


Figure 65: The output of the learning rate finder. The x-axis is the number of trees trained and the y-axis is the auc eval, which is (-auc)

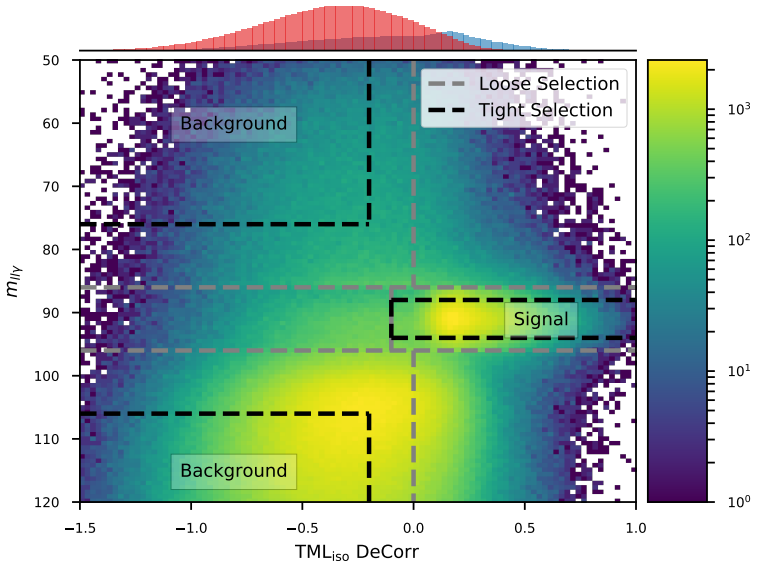


Figure 66: A 2d histogram of the sigmoid decorrelated logit transformed isolation scores and the invariant mass $m_{ll\gamma}$, with the x-axis projected into the 1d histogram, where the loose working point has been used to separate the signal (blue) and background (red) distributions. The separation in this 2d space allows one to select signal and background labels. This is done in two ways, a loose and a tight selection. The latter is a subset of the former. Only the boxes containing a *"Signal"* or *"Background"* label are kept.

select less background in the loose selection. This is most likely not a problem and could be corrected for by loosening the cut.

Acronyms

TT&P Tag, Tag and Probe 47, 52, 78, 79, 85

T&P Tag and Probe 47–49, 78, 79, 81, 93, 105

AUC Area Under Curve 57, 58

BDT Boosted Decision Trees 9, 10, 74–78

CSC Cathode Strip Chambers 37

ECAL Electro-Magnetic Calorimeter 7, 31, 35–37, 39–42, 44, 53

eV Electron volts 21

FNN Feedforward Neural Network 60

HCAL Hadronic Calorimeter 39

HEP High Energy Particle Physics 21, 55

ID Inner Detector 7, 31–33, 39–42, 53

LAr Liquid Argon 35, 36, 38

LeakyReLU Leaky Rectified Linear Unit 60

LGBM LightGBM 59, 64–66, 74, 85, 92

LHC Large Hadron Collider 21, 24, 25, 27, 30, 31, 33

MC Monte-Carlo 8, 9, 12, 13, 45–50, 52, 56, 63–65, 69, 72–74, 76, 82, 84, 92, 94–97, 99, 100, 105

MDT Monitored Drift Tubes 37

MIP Minimally Ionizing Particle 23, 37, 39

MS Muon Systems 32, 37, 39

NI Nuclear Interaction length 24

OS Opposite Sign 48

PDF Parton Distribution Functions 7, 24, 25, 27

PID Particle Identification 78, 93

QCD Quantum Chromodynamics 21, 24, 27

ReLU Rectified Linear Unit 60

RF Random Forest 9, 10, 59, 74–76, 78

ROC Receiver Operating Characteristic 56, 57, 65

RPC Resistive Plate Chambers 37

SCT SemiConductor Tracker 7, 32

SM Standard Model 7, 21, 26, 95

SS Same Sign 48

TGC Thin Gap Champbers 37

TRT Transition Radiation Tracker 7, 32, 33, 39–41